



Improving PPV prediction in open-pit blasting through Cubist-based feature enrichment and machine learning models

Hoang Nguyen, Xuan-Nam Bui, Carsten Drebenstedt & Yosoon Choi

To cite this article: Hoang Nguyen, Xuan-Nam Bui, Carsten Drebenstedt & Yosoon Choi (05 Sep 2025): Improving PPV prediction in open-pit blasting through Cubist-based feature enrichment and machine learning models, International Journal of Mining, Reclamation and Environment, DOI: [10.1080/17480930.2025.2552707](https://doi.org/10.1080/17480930.2025.2552707)

To link to this article: <https://doi.org/10.1080/17480930.2025.2552707>



Published online: 05 Sep 2025.



Submit your article to this journal [↗](#)



Article views: 35



View related articles [↗](#)



View Crossmark data [↗](#)



Improving PPV prediction in open-pit blasting through Cubist-based feature enrichment and machine learning models

Hoang Nguyen^{a,b}, Xuan-Nam Bui^{c,d}, Carsten Drebenstedt^{e,b} and Yosoon Choi^f

^aDepartment of Surface Mining, Mining Faculty, Hanoi University of Mining and Geology, Hanoi, Vietnam;

^bInnovations for Sustainable and Responsible Mining (ISRM) Research Group, Hanoi University of Mining and Geology, Hanoi, Vietnam; ^cLaboratory of Sustainable Development in Natural Resources and Environment, Institute for Advanced Study in Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam; ^dFaculty of Environment and Labour Safety, Ton Duc Thang University, Ho Chi Minh City, Vietnam; ^eInstitute of Mining and Special Engineering, Freiberg University of Mining and Technology, Freiberg, Germany; ^fDepartment of Energy Resources and Geosystems Engineering, College of Engineering, Sejong University, Seoul, Republic of Korea

ABSTRACT

Blast-induced ground vibration poses serious environmental risks in open-pit mines. To improve prediction accuracy, this study introduces a machine learning-based feature enrichment technique using the Cubist algorithm, which generates an additional input variable to enhance the original dataset. Four machine learning models – SVM, RF, k-NN, and GBM – were developed and evaluated on both original and enriched datasets. Results demonstrate that Cubist-based enrichment significantly improves prediction performance, reducing overfitting/underfitting issues. This approach highlights the potential of combining feature enrichment with advanced machine learning models to achieve more accurate and reliable vibration prediction in practical mining applications.

ARTICLE HISTORY

Received 19 August 2024
Accepted 20 August 2025

KEYWORDS

Feature enrichment; mine blasting; blast-induced ground vibration; machine learning; data argumentation; environmental protection

1. Introduction

Blast-induced ground vibration is one of the most significant environmental issues in open-pit mining operations. Excessive vibration can cause structural damage, reduce slope stability or even failure slopes, as well as affect the safety of surrounding communities [1,2]. Accurate prediction of peak particle velocity (PPV) – the most widely used indicator of ground vibration, is therefore essential to minimise environmental risks and optimise blasting designs.

In recent years, a variety of prediction models have been proposed for estimating PPV, as mentioned in Table 1. Traditional empirical formulas often fail to account for the complex, nonlinear relationships among blasting parameters [61]. To overcome these limitations, many researchers have turned to machine learning (ML) techniques such as adaptive neuro-fuzzy inference system (ANFIS), artificial neural network (ANN), support vector machine (SVM), k-nearest neighbour (KNN), classification and regression trees (CART), etc. or even they applied optimisation algorithms to improve the performance of these ML models, such as particle swarm optimisation, genetic algorithm, grey wolf optimisation, harris hawks optimisation, and so on, which can capture hidden patterns and improve prediction accuracy. However, these models still face challenges due to the limited information content of input variables, resulting in reduced model generalisation and interpretability.

To address this gap, the current study introduces a novel feature enrichment strategy using the Cubist algorithm. The main concept is to derive a synthetic variable from the original dataset through a Cubist model, capturing latent nonlinear interactions among the features. This new variable is then incorporated into the original feature set to construct an enriched input space. The resulting dataset, with improved descriptive power, is used to train multiple machine learning models. Their predictive performance is subsequently benchmarked against models trained solely on the original dataset.

Table 1 presented a comprehensive review of previous studies on PPV prediction using various ML/AI techniques. It highlighted the methods, input parameters, and performance metrics used to predict PPV in recent years. From a comprehensive review in Table 1, it can be seen that efforts are being made to improve the accuracy of AI and ML models for predicting PPV in open-pit mines. These efforts aim to address the uncertainties from geological and geophysical conditions, as this field is ever changing. To do this, there are two ways to improve the accuracy of AI and ML models, improving the predictive models (i.e. robustness of the models) and improving the dataset. In recent years, the efforts to improve the robustness of the models are widely studied and applied [62,63]. While many of these studies focus on optimising ML algorithms, none explicitly address the role of input data quality or the potential benefits of feature enrichment. This indicates the originality and novelty of the current research: rather than focusing solely on algorithm optimisation, it explores how enhancing the input dataset itself can lead to better prediction outcomes.

In this study, a novel feature enrichment approach combining data augmentation and machine learning techniques has been proposed to enhance the prediction of blast-induced ground vibration. Initially, a machine learning model is developed to pre-predict PPV from the original dataset, and the predicted outputs are introduced as an additional input variable to encapsulate complex, nonlinear relationships among the existing features. Subsequently, various machine learning models are constructed using both the original and the enriched datasets for comparative evaluation. The novelty and significance of this study lie in the introduction of a data-driven enrichment framework that transcends conventional input configurations. By employing the Cubist algorithm for enrichment, the proposed method addresses limitations in data quality and contributes to improving model generalisation and robustness. This strategy offers a promising direction for advancing predictive modelling in geotechnical and environmental engineering applications.

2. Machine learning approaches

2.1. Machine learning for feature enrichment

As a matter of the fact that machine learning can be a powerful tool for feature enrichment, which involves improving the quality and depth of existing data by adding or enhancing information. In regression problems such as predicting blast-induced ground vibration, machine learning algorithms can be applied to identify outliers, missing values, feature selection and correct errors.

Furthermore, machine learning algorithms can be applied to increase the sample size of existing data, such as Smote (Synthetic Minority Over-sampling Technique), GANs (Generative Adversarial Networks), which can make the original data much more enriching. However, generating synthetic data based on existing data can introduce outliers and impact the accuracy of predictive models, as illustrated in Figure 1.

Another issue of feature enrichment is data augmentation. A simple method for data augmentation is combination of various input variables to generate new input variables based on the addition, subtraction, multiplication, and division calculations. However, such calculations may offer linear relationships between the newly generated input variables and previous ones, and they may affect the performance of predictive models. To deal with this problem, machine learning algorithms are proposed to be used in this study and their feasibility will be evaluated. In fact, machine learning algorithms are often used to provide predictive models based on input variables. In other words,

Table 1. Summary of the AI/ML-based models for predicting PPV (updated based on Nguyen, Bui [3]).

Year	Name of model(s)	Variables used	Performance	Author(s)/References
2008	ANFIS	B, Q, D, E, P_v , VOD	PE = 0%-6%	Singh, Dontha [4]
2009	ANN	HD, B, S, Q, D, BI, E, v , P_v , VOD	MAE = 0.196; R^2 = 0.986	Khandelwal and Singh [5]
2010	MLP	D, Q, B/S, N_{hr} , UCS, T_{dl}	RMSE = 0.031; R^2 = 0.954	Monjezi, Ahmadi [6]
2010	SVM	Q, D	MAE = 0.257; R^2 = 0.960	Khandelwal [7]
2011	GA	HD, B, S, ST, Q, Q_{Σ} , D	PE = 0.028%-2.50%	Verma and Singh [8]
2014	PSO-ANN	d_{hr} , HD, Q, S, B, ST, PF, BI, S_{dr} , N_h	PE = 3.87%-20.14%	Armaghani, Hajihassani [9]
2015	ICA-ANN	B/S, ST, Q, E, P_v , D	R^2 = 0.976	Hajihassani, Armaghani [10]
2015	GEP	N_{hr} , HD, B, S, ST, Q, D_{hr} , D	R^2 = 0.97; MAPE = 4.7%	Dindarloo [11]
2015	DE	HD, B, S, CL, Q, D, E, PF, BI	MAE = 1.92; R^2 = 0.71	Saadat, Hasanzade [12]
2016	ANN-KNN	Q, D	R^2 = 0.658; RMSE = 0.144	Amiri, Amnieh [13]
2016	GBM	H_{hr} , SC, QC, BCF, RS, K, CS_{mjr} , D, Q, PPV	Accuracy = 91.67%	Zhou, Shi [14]
2017	GEP	B, S, ST, HD, d_{hr} , PF, Q, D	R^2 = 0.874; RMSE = 6.732; MAE = 5.164	Faradonbeh and Monjezi [15]
2017	PSO-ANFIS	Q, D	R^2 = 0.940; RMSE = 1.944	Shahnazar, Nikafshan Rad [16]
2017	CART	Q, D	R^2 = 0.95; RMSE = 0.17	Hasanipanah, Faradonbeh [17]
2017	ABC-ANN	Q, D	RMSE = 0.22; MAPE = 4.26%; R^2 = 0.92	Taheri, Hasanipanah [18]
2018	ICA	Q, D	R^2 = 0.94; RMSE = 0.37	Armaghani, Hasanipanah [19]
2018	FCM-SVM	B, S, ST, N_{hr} , Q, D	VAF = 85.25%; RMSE = 1.8; R^2 = 0.853	Sheykhi, Bagherpour [20]
2018	ICA-ANFIS	Q, D	RMSE = 0.220; R^2 = 0.942; VAF = 94.2%	Hasanipanah, Bakhshandeh Amnieh [21]
2019	GA-ANN	Q, D_{hr} , D, SD	R^2 = 0.988; RMSE = 3.047	Azimi, Khoshrou [22]
2019	ICA-SVR, GA-SVR, ABC-SVR, PSO-SVR	Q_{Σ} , D, B, S	For the best model (i.e. GA-SVR): RMSE = 0.267; R^2 = 0.991; MAE = 0.182	Nguyen, Choi [23]
2019	PSO-KNN	Q, D	RMSE = 0.797; R^2 = 0.977	Bui, Jaroopattanapong [24]
2019	HKM-CA	Q_{Σ} , D, H, PF, B, S, ST	RMSE = 0.475; R^2 = 0.995	Nguyen, Bui [25]
2019	MARS	Q, D	R^2 = 0.806; RMSE = 48.867	Hosseini, Tavana [26]
2019	XGBoost	Q_{Σ} , D, H, B, S, ST	R^2 = 0.952; RMSE = 1.742	Nguyen, Bui [27]
2019	BGAMs	Q_{Σ} , D, H, B, S, PF	MAE = 0.430; RMSE = 0.582; R^2 = 0.990	Nguyen, Bui [28]
2020	GA-ANFIS, PSO-ANFIS	B, S, ST, PF, Q, D	For the best model (i.e. GA-ANFIS): R^2 = 0.979; RMSE = 0.240	Yang, Hasanipanah [29]
2020	IM-ANN	B, S, ST, PF, Q, D	R^2 = 0.944; RMSE = 0.378; MAPE = 4.913%	Amiri, Hasanipanah [30]
2020	FFA-ANN	Q, D, S, B, PF	RMSE = 0.464; R^2 = 0.966	Shang, Nguyen [31]
2020	BNN, RF	B/S, HD, ST, PF, Q, D	Accuracy = 90.32%	Zhou, Asteris [32]
2020	HKM-ANN	Q, D, B, PF, S	RMSE = 0.554; R^2 = 0.983	Nguyen, Drebenstedt [33]
2020	BBO-ANN	B, S, ST, PF, Q, RMR, D	R^2 = 0.988; RMSE = 0.026	Li, Kumar [34]
2020	PSO-XGBoost	Q, D, B, PF, S	RMSE = 0.583; R^2 = 0.968; VAF = 96.083	Zhang, Nguyen [35]
2020	FFA-SVR	B, S, ST, Q, D	RMSE = 0.128; R^2 = 0.992	Yang, Rad [36]
2020	FCM-QRNN	Q, D, EI, HD, S, B, PF, ST	RMSE = 0.348; R^2 = 0.961	Bui, Choi [37]
2020	HHO-RF	Q, Q_{Σ} , D_{hr} , D, B, T_{dlr} , f	RMSE = 0.34; R^2 = 0.94	Yu, Shi [38]
2020	HPSOGWO-RVM	Q, Q_{Σ} , D_{hr} , D, B, T_{dlr} , f	RMSE = 0.036; R^2 = 0.986	Yu, Shi [39]
2020	ICA-XGBoost	Q, D, EI, PF, B, S, ST	RMSE = 0.736; R^2 = 0.988	Ding, Nguyen [40]
2020	SaDE-ELM	N_{hr} , Q, D, HD, PF	R = 0.871; MSE = 0.019	Arthur, Temeng [41]
2021	CRANFIS-PSO	B, S, ST, SC, Q, D	RMSE = 0.076; MAPE = 0.624%; R^2 = 0.997	Zhu, Nikafshan Rad [42]
2021	HGS-ANN	Q, T_{dlr} , N_{blast_group} , HD, B, S, ST, PF, D	MAE = 1.153; RMSE = 1.761; R^2 = 0.922	Nguyen and Bui [43]
2021	CSO-ANN	Q, Q_{Σ} , D, S, B, PF	MAE = 0.178; RMSE = 0.246; R^2 = 0.990	Bui, Nguyen [44]

(Continued)

Table 1. (Continued).

Year	Name of model(s)	Variables used	Performance	Author(s)/References
2021	GWO-RVM	Q, B/S, D, ST	MAPE = 15.343; VAF = 83.476	Fattahi and Hasanipanah [45]
2021	SCA-ANN	D, Q, BI, SRH	$R^2 = 0.999$	Lawal, Kwon [46]
2021	JA-XGBoost	D_{hr} , HD, Q, S, BI, B, E, D, DOE, VOD, P_v , PR, CL	RMSE = 4.088; $R^2 = 0.957$	Zhou, Qiu [47]
2021	AGPSO-ELM	B/S, HD, ST, PF, Q, D	$R^2 = 0.900$; MAE = 0.07	Jahed Armaghani, Kumar [48]
2021	AutoencoderNN-SVR	d_{hr} , H, HD, B, S, f, PF, Q, D	RMSE = 1.232; $R^2 = 0.887$	Ke, Nguyen [49]
2021	FFA-BSVR	Q, D	$R^2 = 0.996$; RMSE = 0.065; VAF = 99.83%	Ding, Hasanipanah [50]
2021	WOA-ANN, HHO-ANN, PSO-ANN	Q, N_{blast_group} , T_{dlr} , HD, B, S, ST, PF, D	For the best model (i.e. HHO-ANN): RMSE = 1.537; MAPE = 0.123; $R^2 = 0.930$	Nguyen, Bui [51]
2021	MFO-ANN	HD, Q, N_{hr} , Q_{Σ} , D, RMR	MSE = 0.00077; MAE = 0.022; VAF = 95.983%	Lawal, Kwon [52]
2021	MFO-ANFIS	Q_{Σ} , Q, D, B, S, PF	MAPE = 0.044; $R^2 = 0.986$	Bui, Nguyen [53]
2022	WOA-XGBoost, GWO-XGBoost, BO-XGBoost	d_{hr} , HD, B, S, Q, CL, D, BI, E, PR, P_v , VOD, DOE	For the best model (i.e. WOA-XGBoost): RMSE = 3.054; $R^2 = 0.976$	Qiu, Zhou [54]
2022	FFA-GFFN, ICA-GFFN	ST, B, S, Q_{Σ} , D, Q	For the best model (i.e. FFA-GFNN): RMSE = 0.187; $R^2 = 0.970$	Abbaszadeh Shahri, Pashamohammadi [55]
2022	MARS-PSO-MLP	Q, HD, B, S, ST, PF, D	RMSE = 1.569; $R^2 = 0.902$	Nguyen, Bui [56]
2022	HGS-ANFIS	Q_{Σ} , Q, D, B, S, PF	RMSE = 0.274; $R^2 = 0.985$	Nguyen and Bui [57]
2022	SpaSO-ELM	B, S, f, PF, Q, D	$R^2 = 0.914$; RMSE = 0.389	Nguyen, Bui [58]
2023	MRFO-SONIA	B, S, f, PF, Q, D	MAE = 0.379, MSE = 0.453, MAPE = 0.08, $R^2 = 0.896$	Nguyen, Bui [3]
2023	ANFIS	HD, D, Q	Accuracy = 93%	Yardimci and Erkayaoglu [59]
2023	EO-ANFIS	D, Q	RMSE = 22.314; MAPE = 0.4; $R^2 = 0.744$; VAF = 73.732	Nguyen, Choi [60]

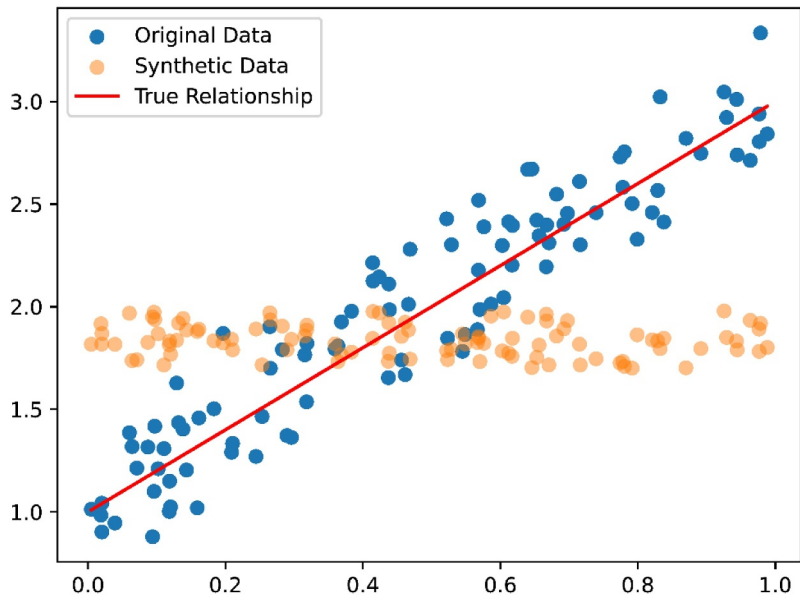


Figure 1. Feature enrichment by the GANs model, aiming to increase sample size of existing data.

predictive models explained the relationships between input and output variables. Most machine learning algorithms are black-box models. Unlike the simple method for data augmentation as mentioned above, machine learning algorithms even can explain non-linear relationships between input and output variables. That's why machine learning was used for feature enrichment, and they will help to create new input variables based on non-linear relationships.

Herein, the Cubist algorithm was applied for feature enrichment of an existing PPV dataset. The Cubist algorithm [64] is a rule-based approach employed for building predictive models through the analysis of input data. In contrast to the See5/C5.0 system introduced by Quinlan [65] for classification tasks, Cubist algorithm excels in solving regression problems. Cubist algorithm's predictive outcomes outperform those of linear regression models and are more straightforward compared to the complex results generated by artificial neural network (ANN) systems [66]. The outcome predictions of the Cubist algorithm then will be used as a new additional input variable for the existing dataset in predicting PPV in this study.

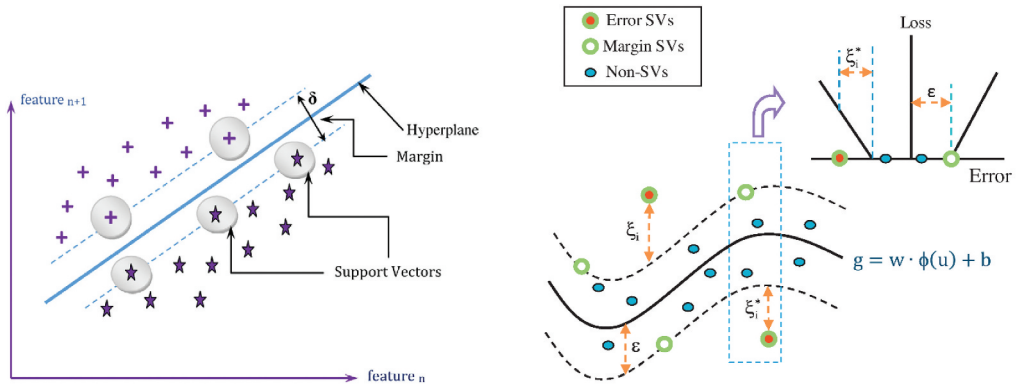
Although the Cubist-based feature enrichment approach can be classified as a black-box method due to the opaque nature of the generated variable, its application in this study is grounded in a clear scientific rationale. The hypothesis is that a synthetic variable – generated through a data-driven model like Cubist – can capture latent, nonlinear relationships among input parameters that may not be fully exploited by conventional features. This enriched variable is expected to enhance the learning capacity of predictive models, particularly those that can benefit from complex, high-dimensional feature spaces. To validate this hypothesis, the enriched dataset was used to train multiple machine learning algorithms and compared with results from models trained on the original dataset.

2.2. Machine learning for predicting PPV

To evaluate the effects of feature enrichment using machine learning (i.e. Cubist algorithm), various machine learning algorithms like Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbors (k-NN), and Gradient Boosting Machine (GBM) were considered to predict PPV on both original and enriched datasets.

SVM is a supervised machine learning algorithm proposed by Cortes and Vapnik [67]. This machine learning algorithm can be used for both regression and classification problems. SVM for regression involves finding a hyperplane that aligns best with the data points so they are as close to the plane as possible. This is the basic idea of SVM regression which is used in both classification and regression problems. To go deeper into SVM, it's important to understand the role of kernel functions. These functions transform non-separable, lower-dimensional data into a higher-dimensional space. This mapping allows the sample points to be linearly separable in this augmented space so SVM can classify or regress the data. When using SVM regression, two key parameters come into play: the penalty parameter (cost, abbreviated as C) and epsilon (ϵ). C plays a crucial role in balancing the smoothness of the decision surface. A smaller C value gives a smoother decision surface which generalises well but might not capture the intricate patterns in the data. A larger C value emphasises the correct classification of all training samples which might result to a decision boundary that fits the training data well. ϵ defines the acceptable margin or distance between the predicted values and the actual values in the training loss function. By tuning these parameters, practitioners can tailor SVM regression models to their datasets, balance accuracy and generalisation. Figure 2 shows the mechanism of the SVM model.

The random forest (RF) algorithm, introduced by Breiman [63], changed regression tree analysis by adding randomness and iteration. Decision trees are great for classifying data based on key features but suffer from overfitting resulting in models with moderate bias and high variance when predicting new samples. These are the same precision (bias) and accuracy (variance) problems in geological, geochemical and blasting data. RF addresses these limitations through two processes. First, each tree is built on a 'bootstrapped dataset' created by



Maximum separation hyperplane

Nonlinear SVR with Vapnik's ε -insensitive loss function

Figure 2. Illustration of the SVM model [49].

randomly sampling with replacement from the training data so that each tree is unique and less sensitive to outliers. Second, 'random feature selection' minimises correlations among trees by forcing them to use a random subset of predictors for nodes so that predictions are more diverse. By combining bootstrapping and random feature selection or 'bagging' RF creates robust models. Evaluation is done internally using an out-of-bag (OOB) subset so no external validation is needed. The aggregation of decision tree predictions depends on the variable type. For regression the average is used, for classification either the statistical mode or a proportion index between categories can be used. RF is simple, can handle large datasets and provides uncertainty metrics but has high computational demands and can't extrapolate beyond the training data range especially in regression modelling. Figure 3 illustrates the mechanism of the RF model.

The k-NN algorithm, a nearest-neighbour approach, is employed for predicting unknown data based on existing datasets. It functions by categorising unknown data based on the closest known data points in empirical knowledge. Comparing unclassified data with existing data, it identifies the categories of the K nearest samples (where K represents the number of samples) and assigns the unclassified data to the majority category among these K samples. This algorithm is versatile, used not only for classification tasks but also for regression predictions. In our case, it's selected for PPV prediction. Figure 4 visually depicts the process of k-NN regression prediction. For regression prediction, the data is initially divided into test and training sets. The algorithm computes weights between the test set and the training set data. It then selects the K training data points with the highest weights to the test set and calculates their average value as the predicted target value.

GBM is a general machine learning algorithm that can be used for regression and classification problems in many fields including geotechnical engineering and mining problems [68–72]. Many studies have shown that GBM outperforms artificial neural networks and other machine learning algorithms, especially in particle identification [73]. GBM improves model accuracy by combining multiple base-learners [74]. Within GBM, two algorithms are used: gradient boosting (GB) and the algorithm to develop base-learners like the decision tree used here. The GBM algorithm is calculated in 5 steps as follows:

Step 1: Import x and y , define the loss function.

Step 2: Create an initial decision tree (DT_0) to estimate the constant that minimises the loss function. This tree defines the function $f(x)$ between x and y .

Step 3: Calculate the pseudo-residuals by taking the derivative of the loss function, use them as the new target. Pseudo-residuals are the residuals since the square error loss function is used here.

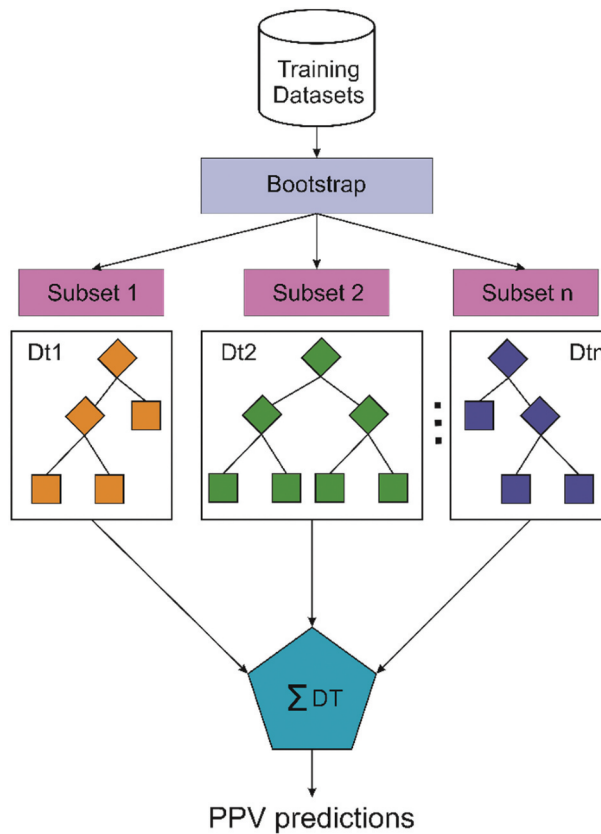


Figure 3. Illustration of the RF model.

Step 4: Create a new decision tree (DT_i) to fit the pseudo-residuals and update $f(x)$ by adding this tree to the model. A linear search is used to find the leaf node value that minimises the loss function.

Steps 3 and 4 are repeated until: a pre-set number of cycles is reached ($i = n_{\text{estimators}}$, chosen through hyper-parameter tuning).

Step 5: Output the GBM model by combining all the trees.

Figure 5 shows the flowchart of the GBM model for a regression problem, like predicting PPV.

3. Study area and data analysis

This study uses the Khanh Hoa open-pit coal mine in Thai Nguyen province, Vietnam as a case study, as shown in Figure 6. This is one of the deepest open-pit coal mines in Vietnam, at -240 m below sea level. The rock hardness in this mine is between 6 to 12.1 on Protodiakonov's scale [75]. To prepare the rock for loading and transportation, blasting methods are used due to the hardness. However, this process is done within 200 to 300 m from surrounding residential areas (as shown in Figure 6) and can affect the nearby communities. So it's important to predict accurately the ground vibrations induced by these blasts to mitigate the environmental impact.

For collecting the dataset, this study considered 8 input parameters: maximum explosive charge per delay (Q), length of borehole (L), stemming (ST), powder factor (q), vertical distance (H), distance between boreholes in a row (a), spacing (S) and monitoring distance of PPV (D). Of these input parameters, the first 7 input parameters were from blast patterns, and the last input parameter was measured by GPS devices. To measure PPV, a Micromate seismograph (manufactured by

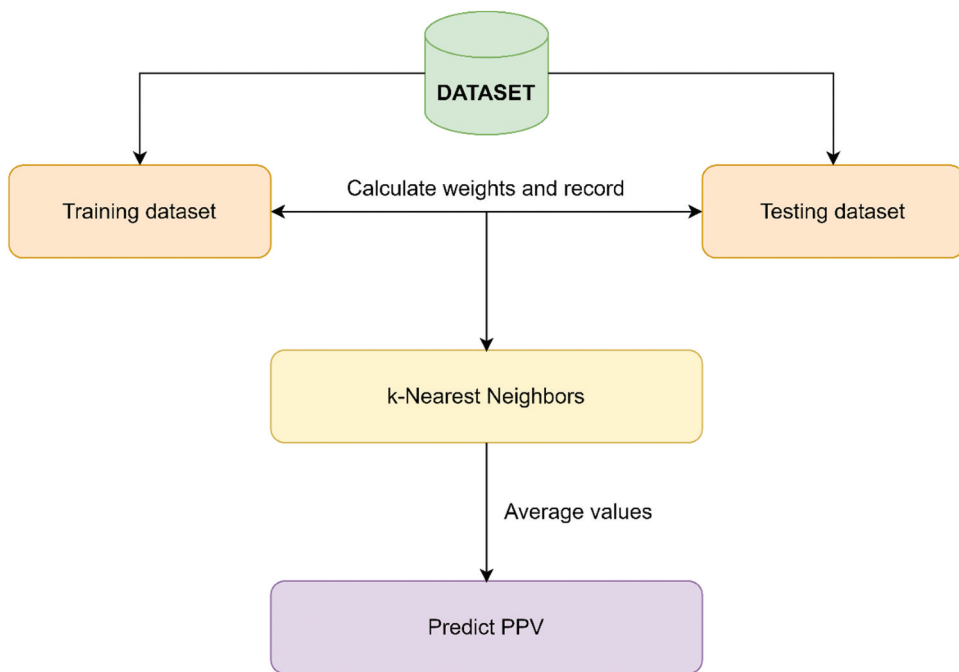


Figure 4. Illustration of the k-NN model for predicting PPV.

Instantel, Canada) was utilised, which is widely used for monitoring blast-induced ground vibrations. This device was installed at varying distances ranging from 57 m to 621 m from the blast sources, depending on the site layout and safety requirements. The Micromate is capable of recording ground vibration velocity in three orthogonal directions (longitudinal, transverse, and vertical), and the Peak Particle Velocity (PPV) was automatically computed by the device based on the maximum vector sum of these components during each blasting event.

All devices were calibrated according to the manufacturer's specifications prior to field deployment to ensure measurement accuracy. Calibration procedures included baseline noise level checks and sensor sensitivity validation. The collected PPV values used in this study represent the maximum particle velocity recorded in any of the three directions during each blast event. Data was logged in real-time and verified by field engineers for consistency and completeness before inclusion in the dataset.

Finally, 183 blasting events were collected and compiled as a dataset to predict PPV at Khanh Hoa open-pit coal mine. The distribution of the collected dataset is shown in [Figure 7](#) and its characteristics are summarised in [Table 2](#). Before developing the predictive models to predict PPV, the dataset needs to be analysed to see the role and contribution of each parameter. So, a correlation matrix was created to see if input variable (s) should be removed from the original dataset since strong correlations as shown in [Figure 8](#).

The heatmap shows pairwise Pearson correlation coefficients between all variables. Most correlations are below 0.75, suggesting low multicollinearity among input variables. A few notable relationships are seen between L and Q ($r = 0.72$), and between L and ST ($r = 0.70$), indicating some dependency among design parameters. Importantly, no input variable shows a strong correlation with the output PPV, suggesting a complex, nonlinear relationship that supports the use of machine learning models for prediction. So there is no input variable to be removed in this study. In other words, all 8 input variables were used for feature enrichment and model development in this study.

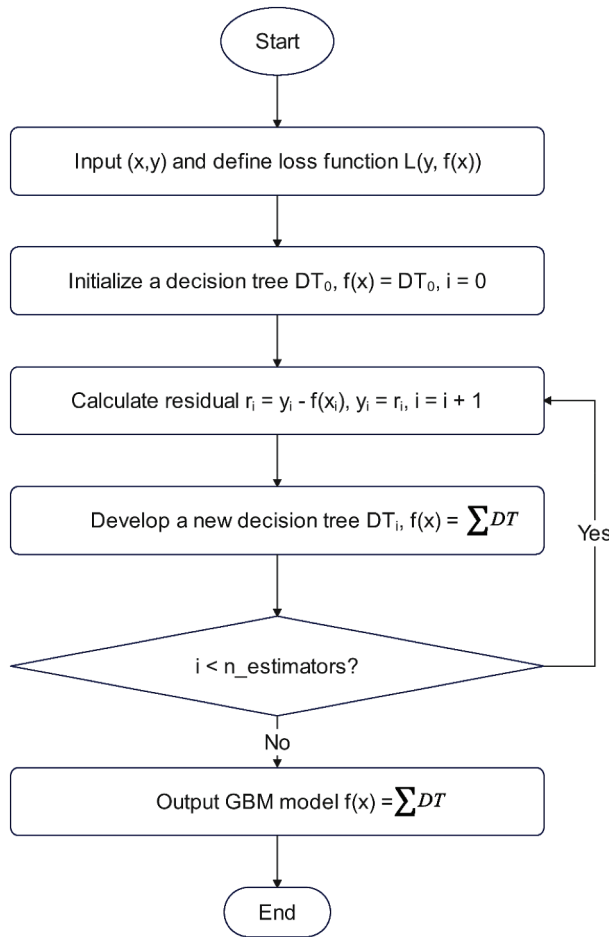


Figure 5. Illustration of the GBM model.

4. Results and discussion

Before developing the PPV models in this study, the dataset was split into 70% for training and 30% for testing to ensure that the machine learning models had sufficient data to learn robust patterns while preserving a separate dataset for evaluating generalisation performance. This 70/30 split is widely adopted in predictive modelling studies to balance model learning and performance assessment. Furthermore, a 10-fold cross-validation procedure was applied to the training data to assess model stability and reduce the risk of overfitting. While 5-fold cross-validation is also commonly used, 10-fold cross-validation provides a more comprehensive evaluation by reducing variance in performance estimates, especially for moderate-sized datasets. RMSE was used as the loss function to train the PPV models.

Then, the Cubist algorithm was applied to build a model on the original dataset with number of instances and number of committees to control the accuracy of the Cubist model. In fact, the Cubist model is a PPV model built based on 8 selected input variables as described above. In other words, the Cubist model explained the relationship between the input variables as shown in Figure 9.

By this way, the Cubist model output was added to the original dataset as a new input variable that describes the relationship between the original input variables through the f_{Cubist}

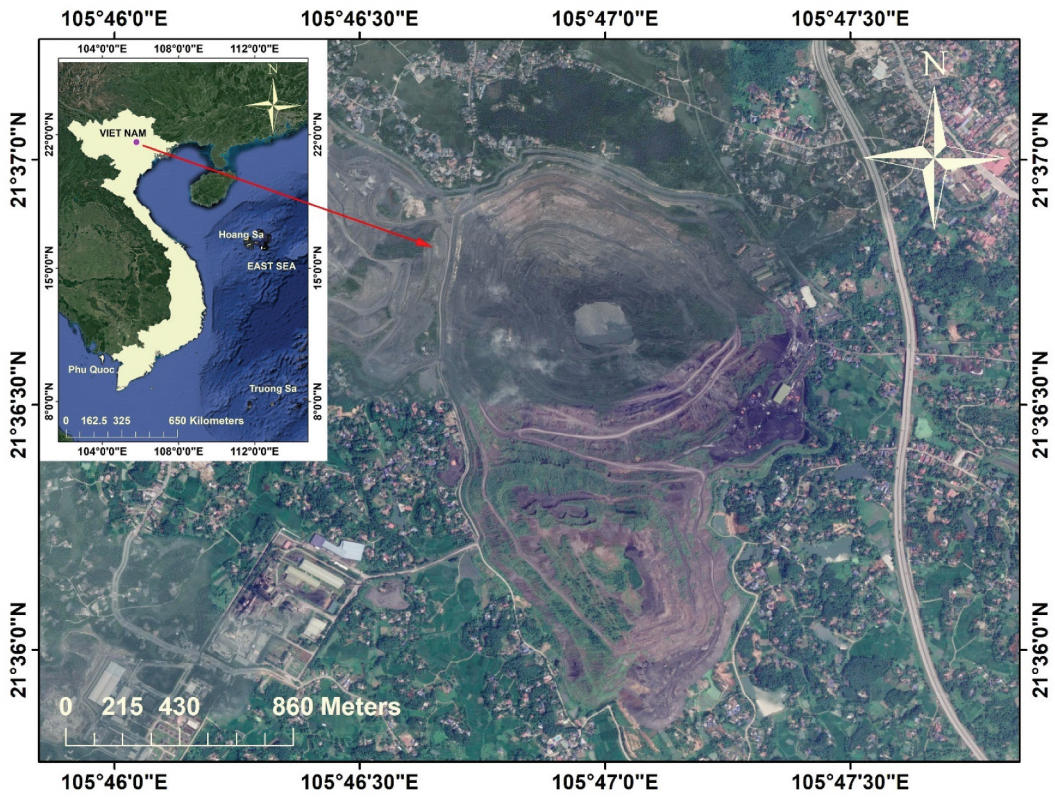


Figure 6. Location of the Khanh Hoa open-pit coal mine.

Table 2. Summary of the collected dataset (original dataset).

Statistics	Q (Kg)	L (m)	ST (m)	q (kg/m ³)	D (m)
Min.	12.00	3.60	0.298	0.1890	57.0
1st Qu.	60.00	8.05	5.100	0.2935	145.5
Median	78.00	10.20	5.700	0.3150	234.0
Mean	87.13	10.90	6.013	0.3603	250.5
3rd Qu.	128.00	13.65	6.650	0.3500	322.0
Max.	186.00	20.10	9.600	5.9000	621.0
Statistics	H (m)	a (m)	S (m)	PPV (mm/s)	–
Min.	1.00	2.500	2.500	0.190	–
1st Qu.	19.00	4.500	4.000	2.080	–
Median	41.00	5.000	4.900	4.270	–
Mean	59.66	4.847	4.566	6.561	–
3rd Qu.	100.00	5.150	5.000	8.745	–
Max.	220.00	5.700	5.200	30.700	–

function. A new dataset (enriched features) with 9 input variables has been generated. The additional input variable distribution is shown in Figure 10 and the enriched dataset properties are in Table 3.

To assess the potential redundancy or complementarity of the Cubist-generated variable, a correlation matrix was constructed for the enriched dataset, as shown in Figure 11. The analysis indicated that the new variable exhibited moderate correlation with some original inputs (maximum $|r| < 0.75$), suggesting that it captures nonlinear relationships not directly encoded in the raw

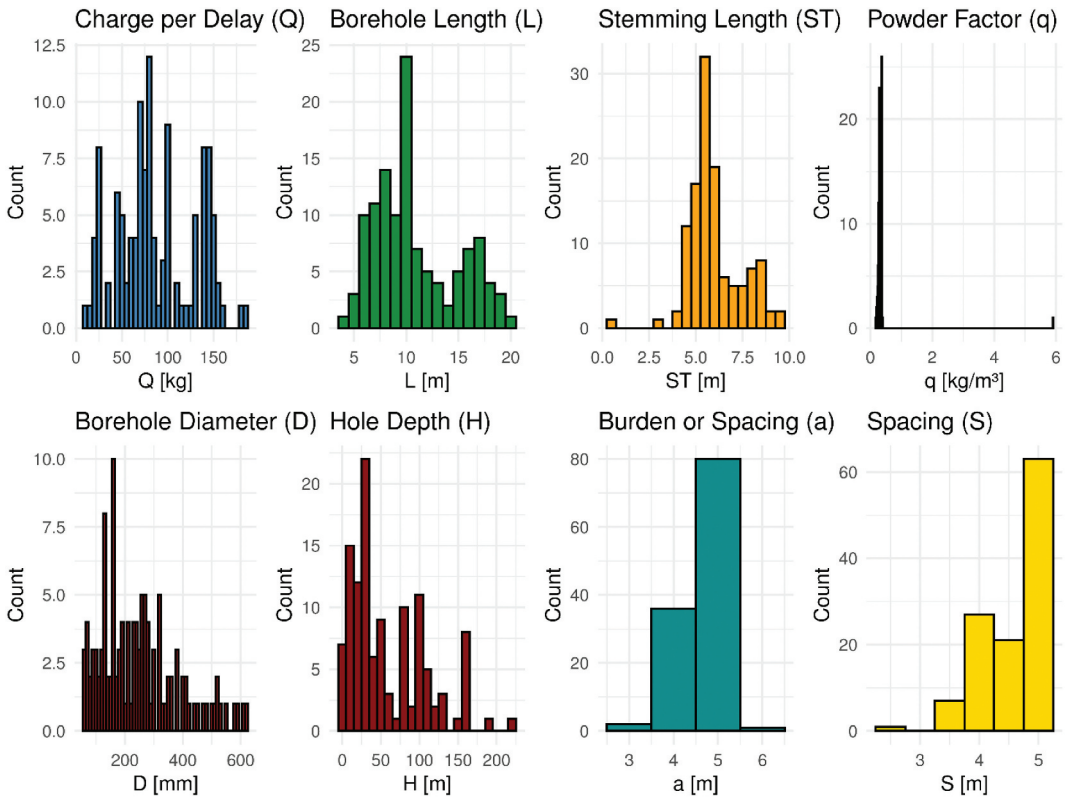


Figure 7. Histograms showing the distribution of the eight input parameters used in this study: maximum explosive charge per delay (Q, kg), borehole length (L, m), stemming length (ST, m), powder factor (q, kg/m³), vertical distance (H, m), distance between boreholes in a row (a, m), spacing between rows (S, m), and monitoring distance of PPV (D, mm). Each parameter is plotted separately to preserve its unit and scale, ensuring accurate interpretation of the input data.

features. This reinforces its value as an informative addition to the feature space, justifying its integration in the model development process.

After feature enrichment with Cubist and new dataset creation, four machine learning models were developed for PPV (SVM, RF, k-NN, GBM) on both the original and enriched datasets and compared them to see the effect of feature enrichment with Cubist.

For the SVM model, the RBF kernel was utilised and controlled its accuracy with the cost (C) and sigma. RMSE was used as the loss function with 3 repeated 10-fold cross-validation to avoid overfitting. The grid search technique was applied to tune the hyperparameters of the SVM model with C between 30 to 60 and sigma between 0.01 to 0.04. The performance of the SVM model on the original and enriched training datasets is shown in Figure 12. The best parameters for the SVM on the original training dataset is $C = 55$ and $\sigma = 0.02$ and on the enriched dataset is $C = 41$ and $\sigma = 0.01$.

For the RF model, its accuracy was controlled with the number of trees in the forest and the number of random input variables. To ensure the diversity of the forest, we created various numbers of trees: 500, 1000, 1500 and 2000. For each number of trees, an RF model was created with different random input variables. Four RF models were created for each dataset (original and enriched) as shown in Figure 13. The results showed that 2000 trees gave the best performance for the RF model on both original and enriched training datasets with 8 and 9 random input variables respectively.

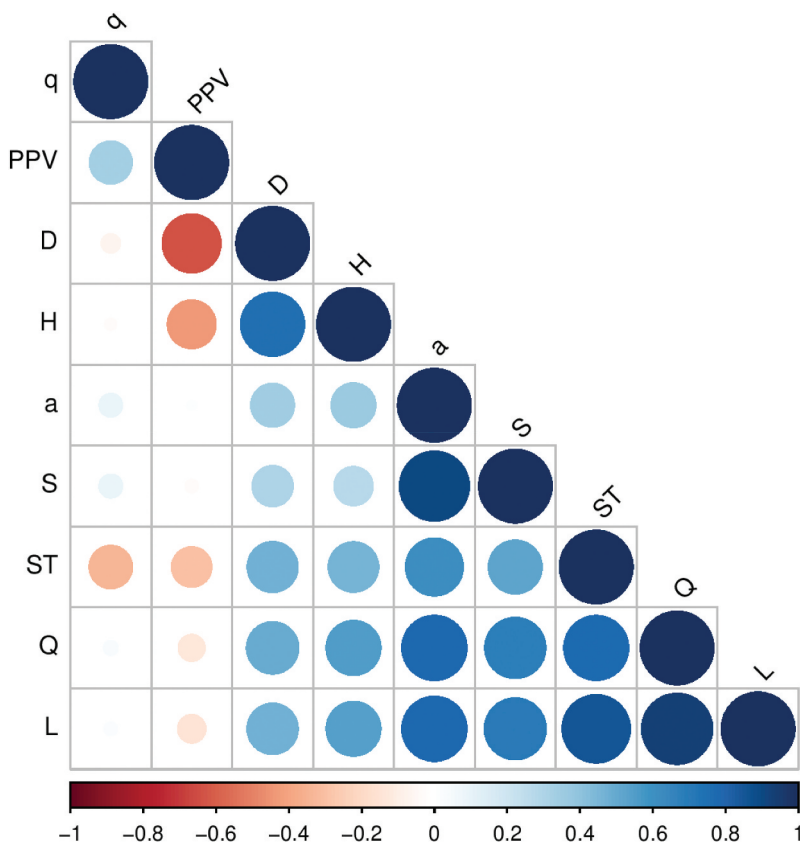


Figure 8. Correlation matrix of the original dataset.

During the development of the k-NN model for PPV in this study, the maximum neighbours, distance and various kernel functions were used to fine tune the k-NN model. Like SVM and RF models, 3 repeated 10-fold CV with grid search were utilised to avoid overfitting of the model. The performance of the k-NN models on the original and enriched training datasets is shown in Figure 14. The best parameters of the k-NN models are maximum neighbours = 6, distance = 5, kernel = biweight on the original training dataset and maximum neighbours = 6, distance = 1, kernel = gaussian on the enriched training dataset.

The GBM model is controlled by parameters such as shrinkage, iteration depth, min observation in node and number of trees. Random search technique with 10-fold cross-validation was applied to find the best GBM models for PPV on both original and enriched training datasets. Several representative examples and model configuration details are presented in Tables 4 and 5, while the full results are provided in the appendix (Tables A1 and A2).

Once the machine learning models were trained for PPV, the test datasets on original and enriched datasets were applied to evaluate the performance of the trained models. Then they were compared on both original and enriched datasets as shown in Table 6.

Upon evaluating the performance of the developed machine learning models on original dataset, it is clear that SVM and RF models do not show overfitting or underfitting. k-NN model is overfitting and GBM model is slightly underfitting. So SVM model is the best fit for original dataset with RMSE = 3.805, $R^2 = 0.714$, MAE = 2.825 on training dataset and RMSE = 3.150, $R^2 = 0.747$, MAE = 2.434 on test dataset.

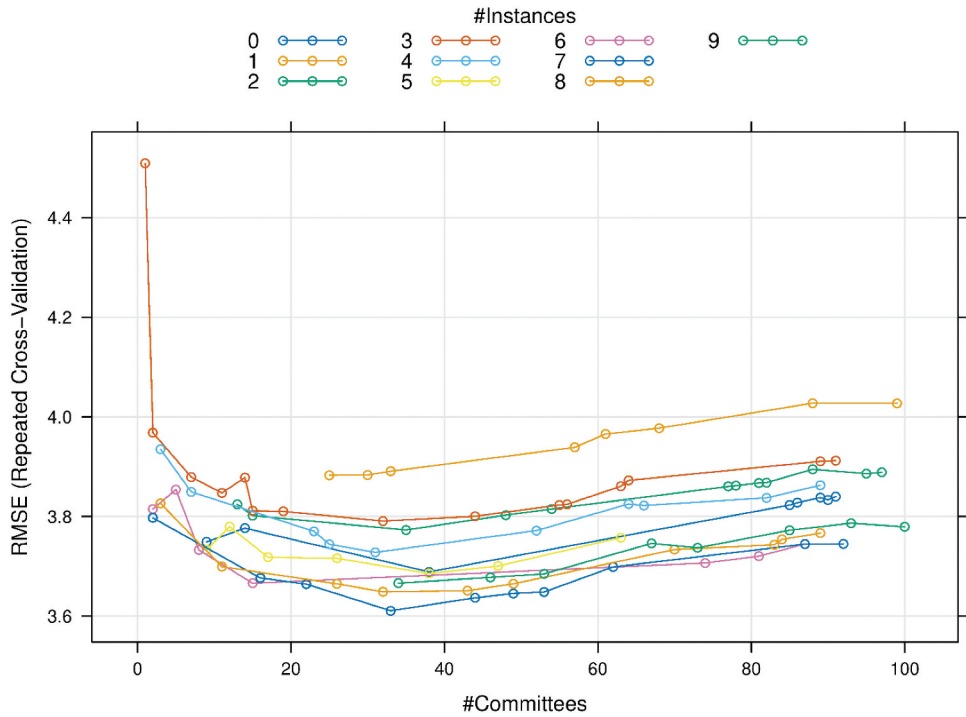


Figure 9. Cubist model for feature enrichment in this study with RMSE.

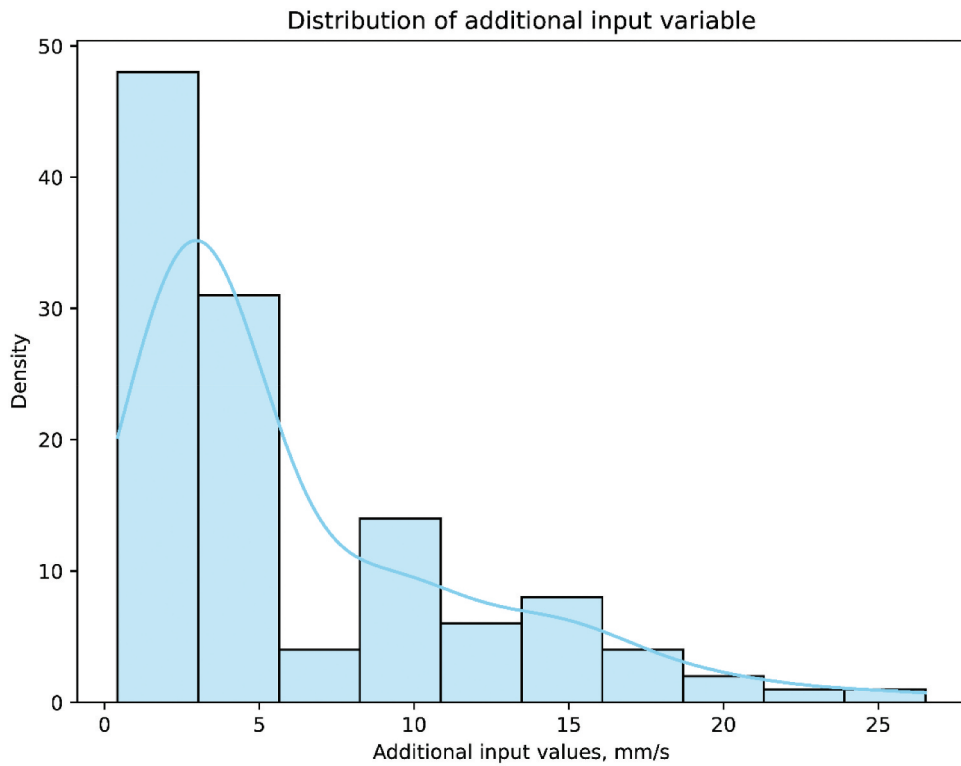


Figure 10. Distribution of the new additional input variable Cubist.

Table 3. Summary of the enriched dataset.

Statistics	Q	L	ST	q	D
Min.	12.00	3.60	0.298	0.1890	57.0
1st Qu.	60.00	8.05	5.100	0.2935	145.5
Median	78.00	10.20	5.700	0.3150	234.0
Mean	87.13	10.90	6.013	0.3603	250.5
3rd Qu.	128.00	13.65	6.650	0.3500	322.0
Max.	186.00	20.10	9.600	5.9000	621.0
Statistics	H	a	S	Cubist	PPV
Min.	1.00	2.500	2.500	0.4195	0.190
1st Qu.	19.00	4.500	4.000	2.5283	2.080
Median	41.00	5.000	4.900	3.7564	4.270
Mean	59.66	4.847	4.566	6.2271	6.561
3rd Qu.	100.00	5.150	5.000	8.9707	8.745
Max.	220.00	5.700	5.200	26.5195	30.700

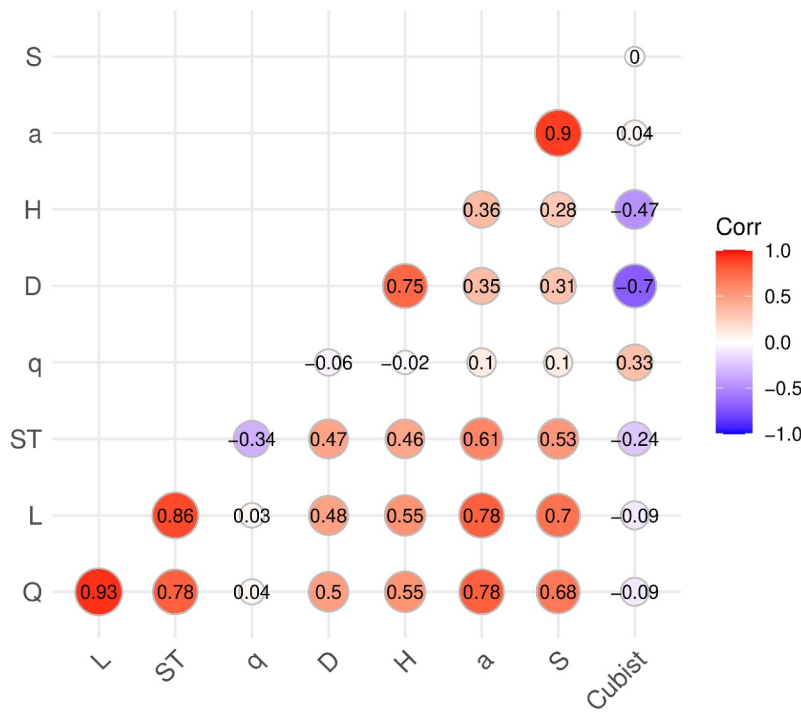


Figure 11. A correlation matrix of the enriched dataset.

However upon evaluating the performance of these machine learning models on enriched dataset, all four models are performing better without any overfitting or underfitting. In other words enriched dataset is more aligned with these machine learning models, not only improving their performance but also fixing the overfitting or underfitting issues in original dataset. Compared to the models developed on original dataset, SVM model is improved 2%, RF model is improved 15%, k-NN model is improved 13% and fixed the overfitting problem and GBM model is improved 15% and fixed the underfitting problem on enriched dataset. On enriched dataset RF is the best model for PPV with RMSE = 2.415, $R^2 = 0.874$, MAE = 1.775 on training dataset and RMSE = 2.611, $R^2 = 0.826$, MAE = 2.083 on test dataset.

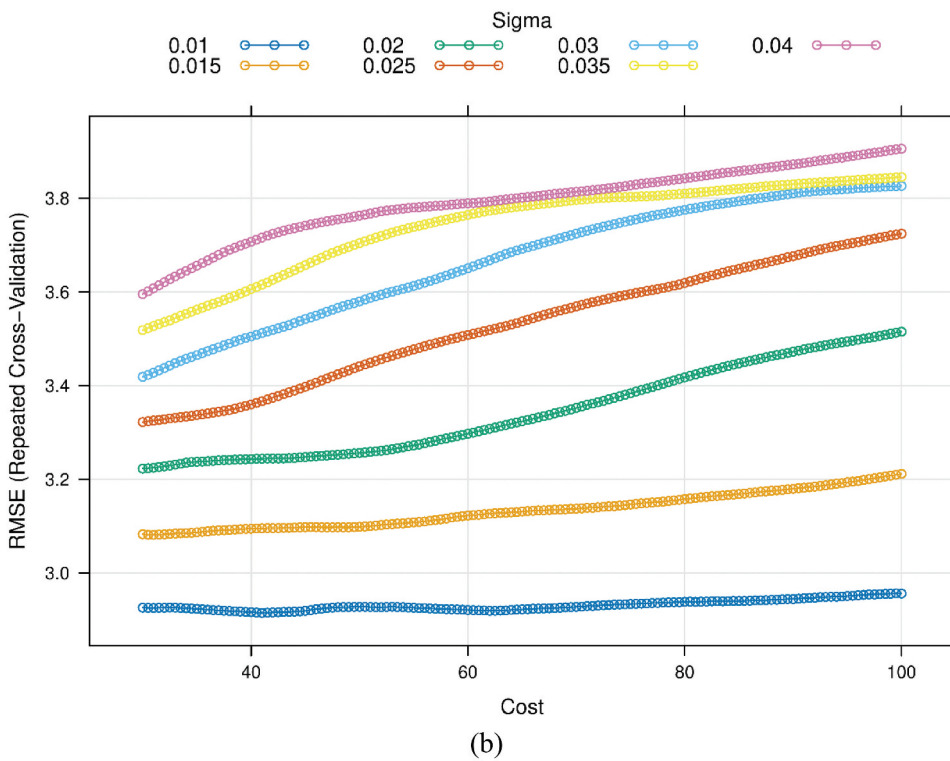
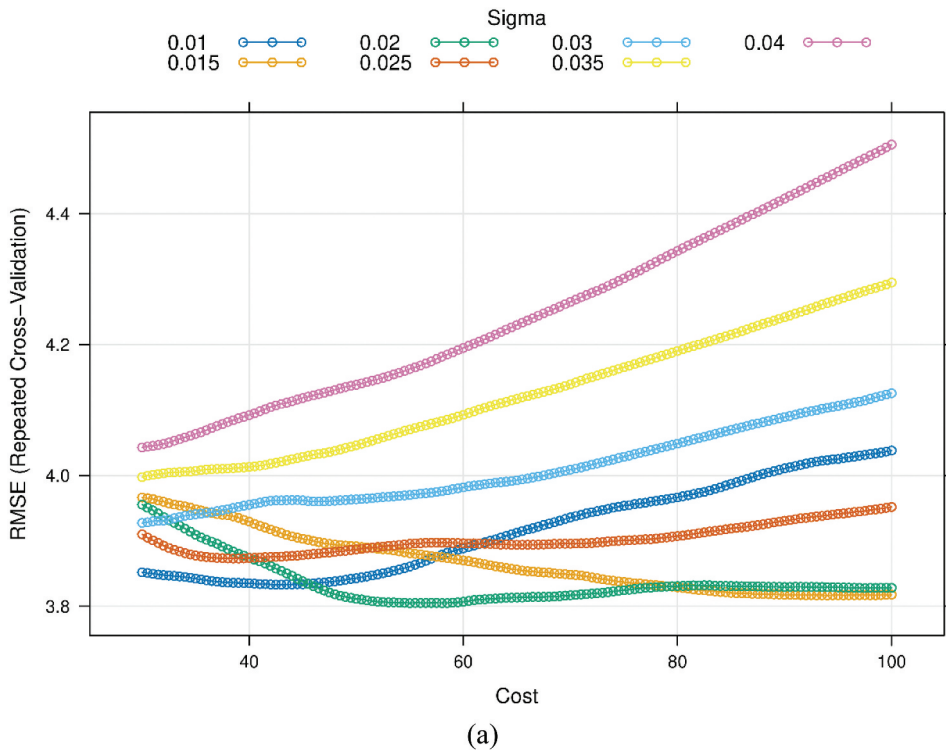


Figure 12. Training performance of the SVM model on the original and enriched datasets (a) on the original dataset; (b) on the enriched dataset.

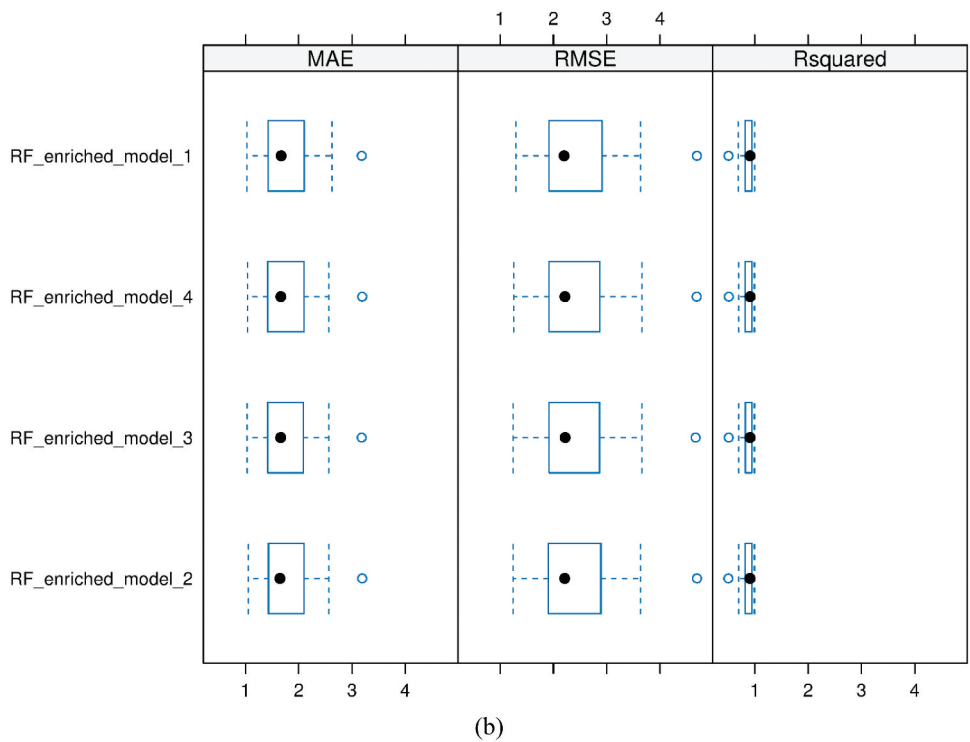
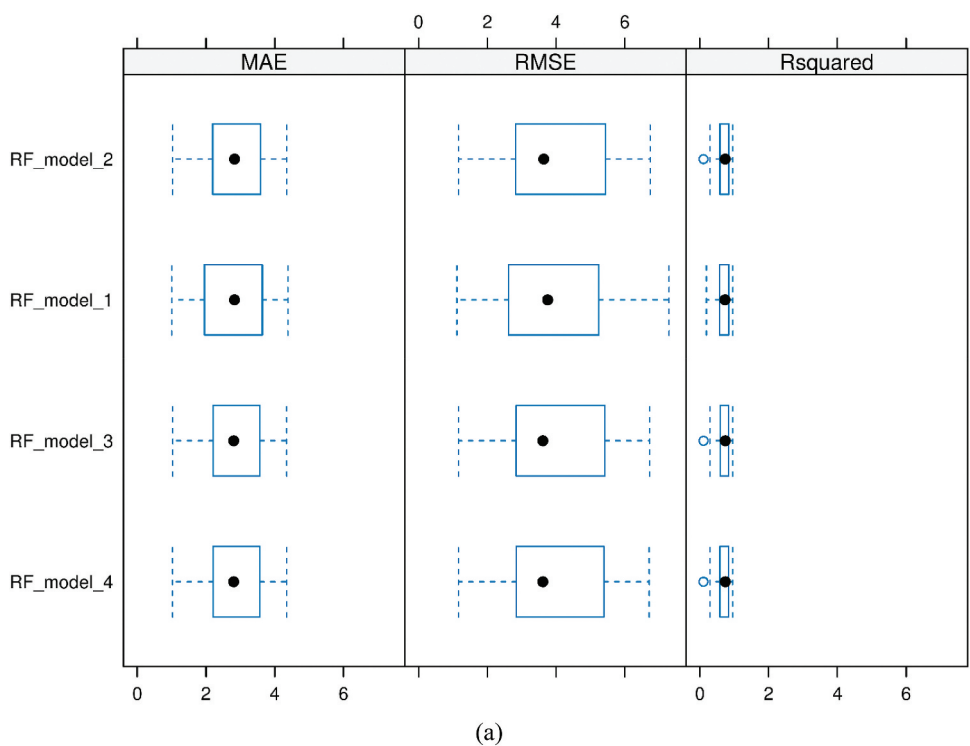


Figure 13. Training performance of the RF model on the original and enriched datasets (a) on the original dataset; (b) on the enriched dataset.

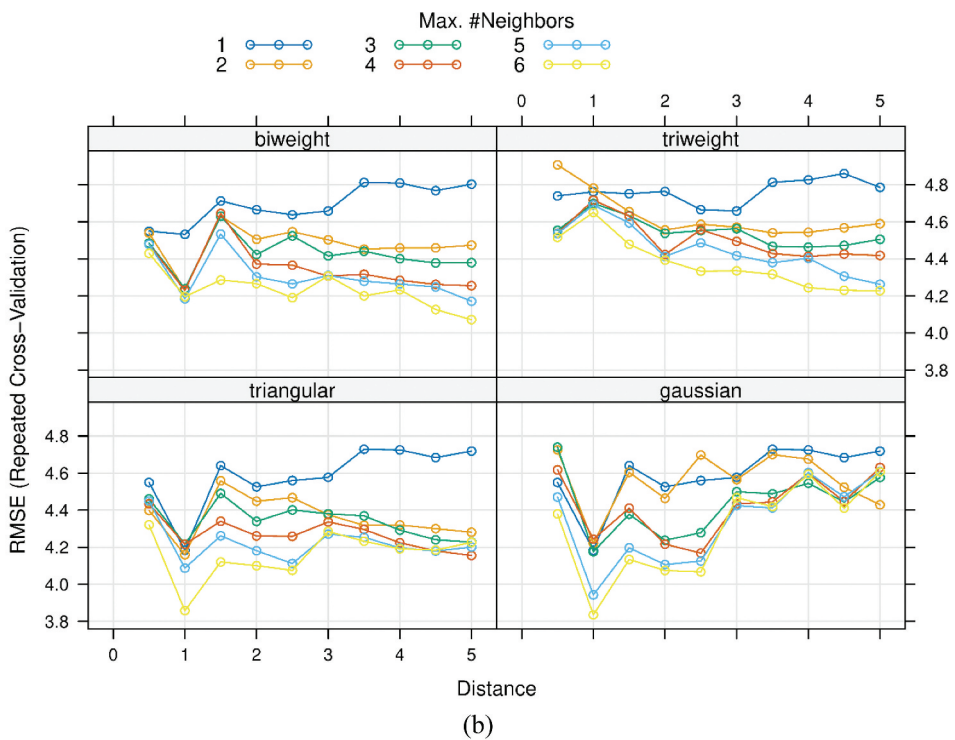
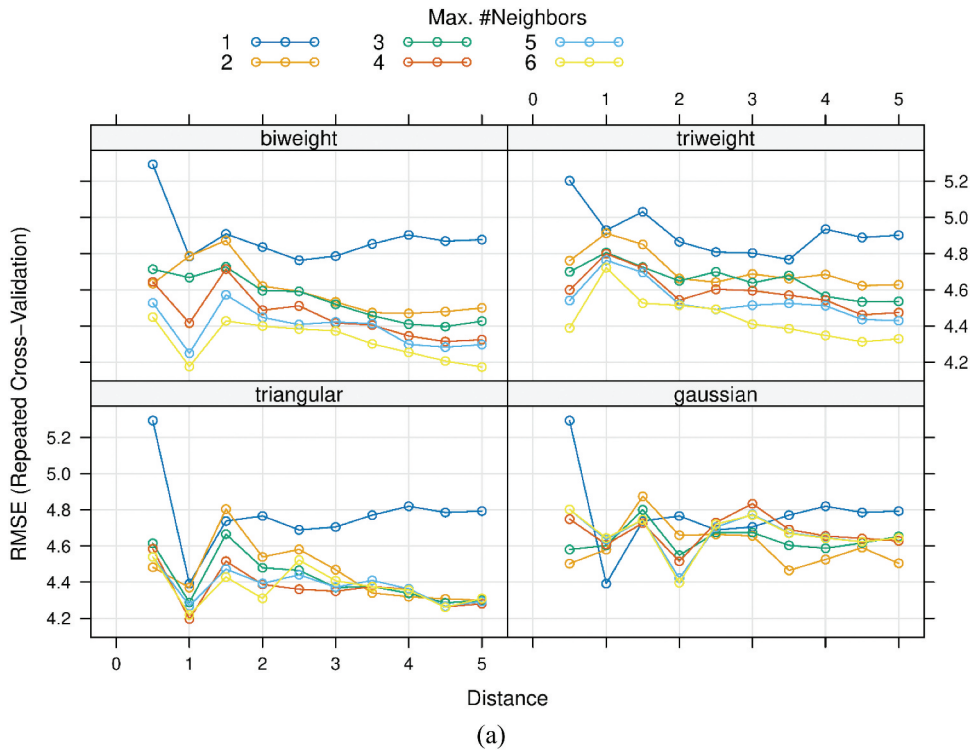


Figure 14. Training performance of the k-NN model on the original and enriched datasets (a) on the original dataset; (b) on the enriched dataset.

Table 4. Training performance of the GBM model on the original dataset.

No.	shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	R ²	MAE
1	0.001149	4	22	2351	NA	NA	NA
2	0.014694	4	13	3674	4.386575	0.596339	3.276882
3	0.015137	2	9	1375	4.357478	0.614967	3.103433
4	0.022373	7	21	1706	NA	NA	NA
5	0.025581	10	25	2057	NA	NA	NA
6	0.038364	9	5	270	3.807603	0.721686	2.642149
7	0.043673	7	17	2316	4.501535	0.597116	3.436295
8	0.046366	3	25	3254	NA	NA	NA
...	...	...	...	...	...	...	...
98	0.586788	3	6	141	4.562059	0.650487	3.267443
99	0.595799	9	12	3018	6.102514	0.37856	4.737745
100	0.596342	1	8	625	5.451697	0.462033	4.084655

Note: NA is not a number. The best configuration is in bold type.

Table 5. Training performance of the GBM model on the enriched dataset.

No.	shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	R ²	MAE
1	0.001149	4	22	2351	NA	NA	NA
2	0.014694	4	13	3674	4.132822	0.676041	2.964837
3	0.015137	2	9	1375	3.750214	0.738641	2.593016
4	0.022373	7	21	1706	NA	NA	NA
5	0.025581	10	25	2057	NA	NA	NA
6	0.038364	9	5	270	2.907844	0.835758	2.058606
7	0.043673	7	17	2316	4.39439	0.621619	3.349479
8	0.046366	3	25	3254	NA	NA	NA
...	...	...	...	...	...	...	...
98	0.586788	3	6	141	3.939274	0.678588	2.870145
99	0.595799	9	12	3018	5.495522	0.499076	4.27109
100	0.596342	1	8	625	4.556901	0.604601	3.437141

Note: NA is not a number. Note: NA is not a number. The best configuration is in bold type.

Table 6. Comparison of the models' performance on the original and enriched datasets.

Machine learning models	Training performance			Testing performance		
	RMSE	R ²	MAE	RMSE	R ²	MAE
Original dataset						
SVM	3.805	0.714	2.825	3.150	0.747	2.434
RF	4.054	0.694	2.834	3.519	0.684	2.415
k-NN	4.173	0.605	2.992	2.865	0.791	2.270
GBM	3.808	0.722	2.642	3.587	0.672	2.327
Enriched dataset						
SVM	2.915	0.816	2.189	3.013	0.768	2.213
RF	2.415	0.874	1.775	2.611	0.826	2.083
k-NN	3.835	0.731	2.725	3.320	0.719	2.591
GBM	2.908	0.836	2.059	2.664	0.819	2.098

A comparative analysis of model performance on the original and enriched datasets (Table 6) demonstrates that the proposed feature enrichment approach yielded substantial benefits for models with lower initial predictive capacity. Specifically, the R² values of RF and GBM increased notably from 0.684 to 0.826 and from 0.672 to 0.819, respectively. A modest improvement was also observed for the SVM model (from 0.747 to 0.768). In contrast, the performance of the k-NN model declined slightly (from 0.791 to 0.719). These outcomes suggest that the effectiveness of feature enrichment is model-dependent. Tree-based algorithms like RF and GBM, which can flexibly capture nonlinear interactions, appear to benefit from the additional explanatory power provided

by the synthetic feature. On the other hand, distance-based algorithms such as k-NN may be adversely affected by increased feature dimensionality or redundant information, which can distort similarity measures. Therefore, while feature enrichment is a valuable strategy for improving prediction in models with limited original performance, its application should be evaluated carefully for models that are already well-tuned or sensitive to feature space alterations.

The comparative analysis in Table 6 demonstrates that the enrichment process improved predictive accuracy for some models, notably RF and GBM, while having limited or negative effects on others like k-NN. These results are not random but reflect how each algorithm interacts with the enriched variable. Tree-based models such as RF and GBM are inherently capable of leveraging complex, nonlinear features, and thus benefit from the additional explanatory power provided by the synthetic variable. In contrast, the k-NN model, which relies on distance metrics, may suffer from added noise or distorted feature space due to high feature correlation or scaling issues. These findings confirm the hypothesis that feature enrichment is more suitable for models with limited capacity to capture nonlinearity in the original feature set, and less effective – or even detrimental – for algorithms that are already optimised for the original space. Therefore, algorithm-specific considerations are crucial when designing enrichment strategies.

To further illustrate the effects of feature enrichment, a comparative bar chart of RMSE, MAE, and R^2 values for all developed models on both original and enriched datasets was constructed (Figure 15). The results clearly show that all four machine learning models benefit from feature enrichment through the Cubist algorithm. Notably, the RF model demonstrated the most significant improvement, achieving the lowest RMSE (2.415), lowest MAE (1.775), and highest R^2 (0.874) on the enriched dataset. This highlights its superior predictive accuracy and consistency. GBM and SVM also exhibited meaningful performance gains, while k-NN showed moderate improvements. The consistent reduction in RMSE and MAE, coupled with an increase in R^2 across all models,

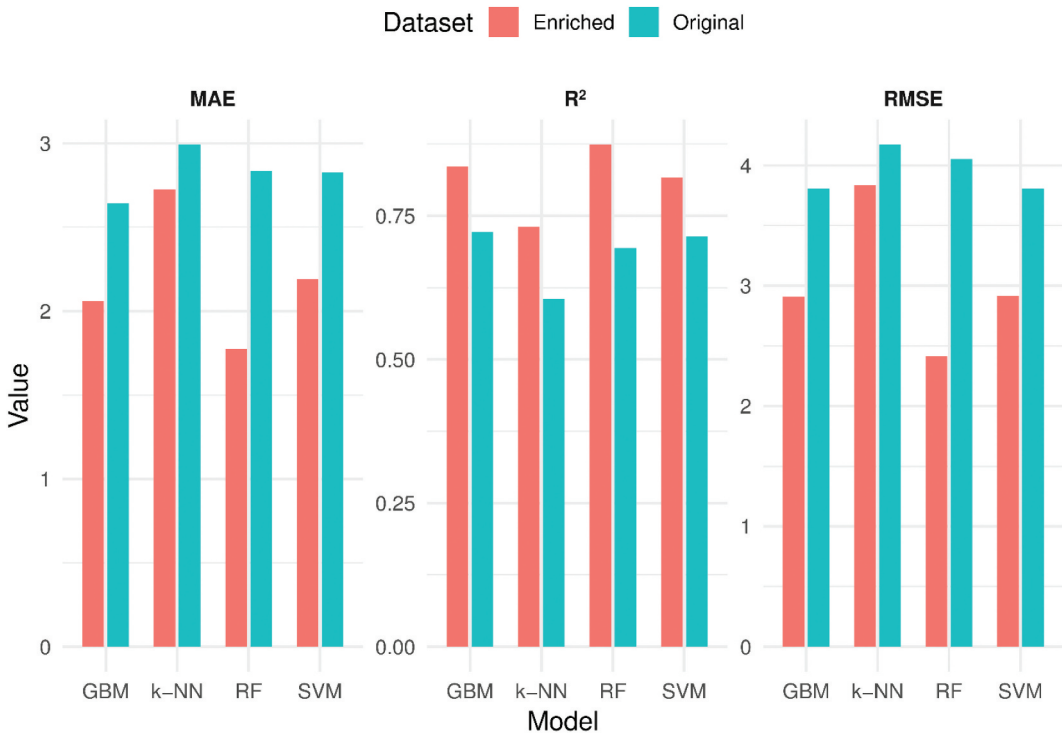


Figure 15. Bar chart comparing RMSE, MAE, and R^2 of models on original and enriched datasets.

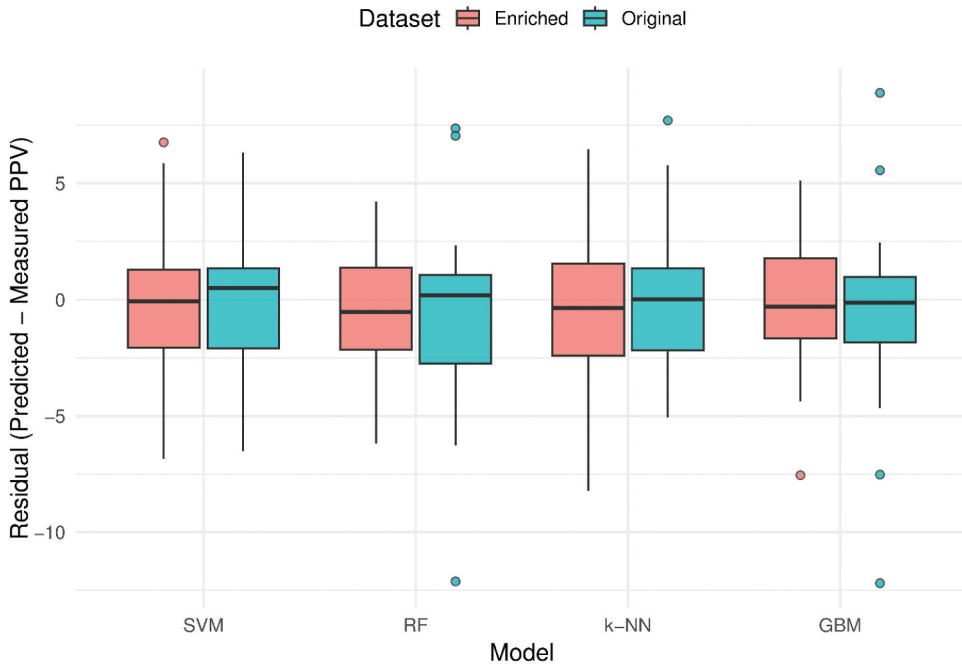


Figure 16. Boxplots of residuals for each model on both datasets.

confirms that the enriched dataset better represents the underlying relationships and reduces both systematic and random prediction errors. These findings support the proposed feature enrichment framework as a viable strategy to enhance machine learning-based prediction of blast-induced ground vibration in open-pit mines.

Figure 16 presents the distribution of residuals (i.e. predicted minus measured PPV values) for each machine learning model on both the original and enriched datasets. As shown, all models exhibit improved performance on the enriched dataset, evidenced by narrower interquartile ranges and fewer outliers. The RF and GBM models display the most compact residual distributions after enrichment, indicating higher prediction consistency and reduced variance. Moreover, the medians of residuals across all models on the enriched dataset are closer to zero, suggesting reduced systematic bias. In contrast, the models trained on the original dataset reveal wider residual spread and more extreme values, particularly for k-NN and GBM, reflecting higher uncertainty and occasional large prediction errors. These findings confirm that feature enrichment significantly enhances both the accuracy and robustness of predictive models for blast-induced ground vibration in open-pit mining.

Figure 17 presents the residual plots for the four machine learning models – SVM, RF, k-NN, and GBM – on both the original and enriched datasets. These plots illustrate the behaviour and distribution of residuals in relation to measured PPV values, offering insight into model bias and consistency. Across all models, the residuals on the enriched datasets show tighter and more uniform dispersion around the zero line, indicating reduced bias and improved generalisation. Notably, the RF model on the enriched dataset demonstrates the most balanced and random residual pattern, with minimal clustering and no evident trends – signifying an excellent model fit. In contrast, the residuals from the original dataset exhibit more pronounced scatter and heteroscedasticity, especially for the k-NN and GBM models, where residual variance increases with higher PPV values. This suggests that these models struggled to accurately capture extreme vibration behaviours when trained only on the original dataset. Overall, the enriched dataset leads

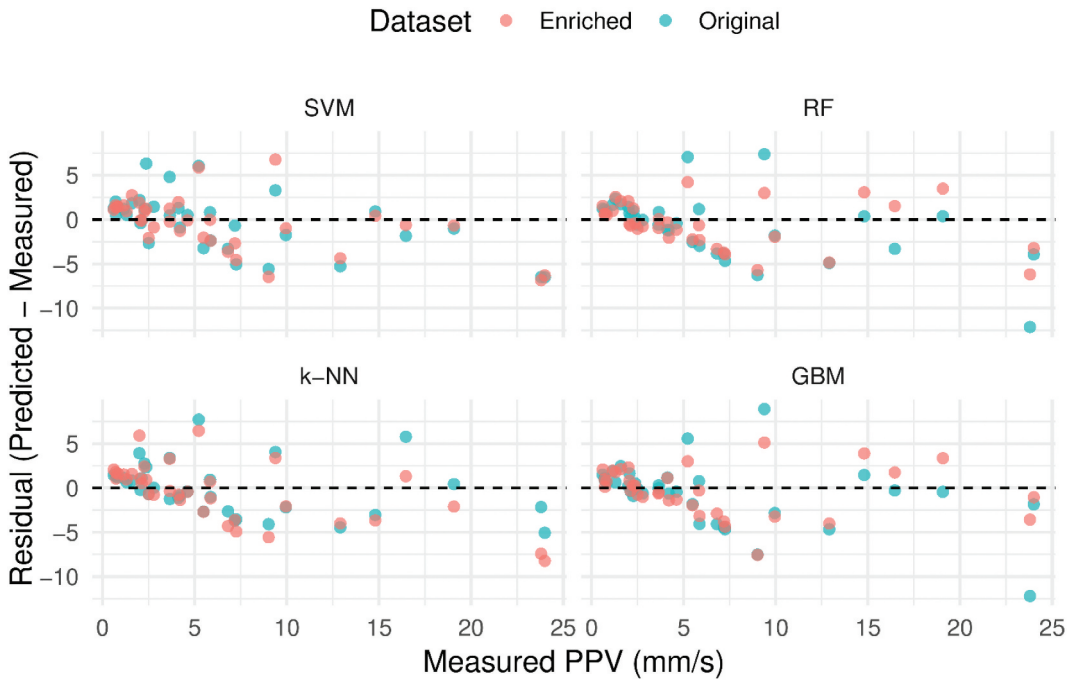


Figure 17. Residual plots (Predicted – Measured PPV vs. Measured PPV) for each model.

to visibly improved residual structure across all models, confirming that the Cubist-based enrichment approach enhances the predictive stability and reliability of machine learning models in modelling blast-induced ground vibrations.

Figure 18 shows the regression analysis of the developed models on both original and enriched datasets. This figure illustrates the improvement of machine learning models when applied to enriched data for blast induced ground vibration.

As shown in Figure 18, it can be seen that the machine learning models are converging better on enriched dataset than on original dataset, especially for RF and GBM model. This shows that Cubist algorithm played a significant role in improving the accuracy of machine learning models for PPV in open-pit mines. Although it is a black-box model; however, it provided more details of the dataset as well as the relationships between the original input variables. This helped SVM, RF, k-NN and GBM models to understand the dataset better and predict more accurately. To further demonstrate the importance of Cubist algorithm in improving the accuracy of PPV predictive models, the best model, RF, on enriched dataset was selected to analyse the variable importance. The variable importance values from IncNodePurity measure in a Random Forest model shows how much each variable contributes to the model's ability to make accurate predictions. In this case, the higher the IncNodePurity value for a variable, the more important that variable is in making decisions in the Random Forest model for PPV. The results are shown in Figure 19.

As shown in Figure 19, it is clear that Cubist variable (additional input variable) has the highest importance among the variables used to predict PPV in enriched dataset with the value of 3270.1466. Its high IncNodePurity shows it plays a big role in the model's predictions. Its magnitude is huge so it significantly affects the PPV predictions. 'D' also has high importance value (52.09429) so it strongly affects the model's decisions. 'H' is another important variable, though not as important as 'D'. It still has significant weight in the model with importance value of 31.74676. 'q' has high importance value (29.35206) so it is one of the key features affecting the model's predictions. 'q' and 'ST' have similar importance values so they contribute significantly to

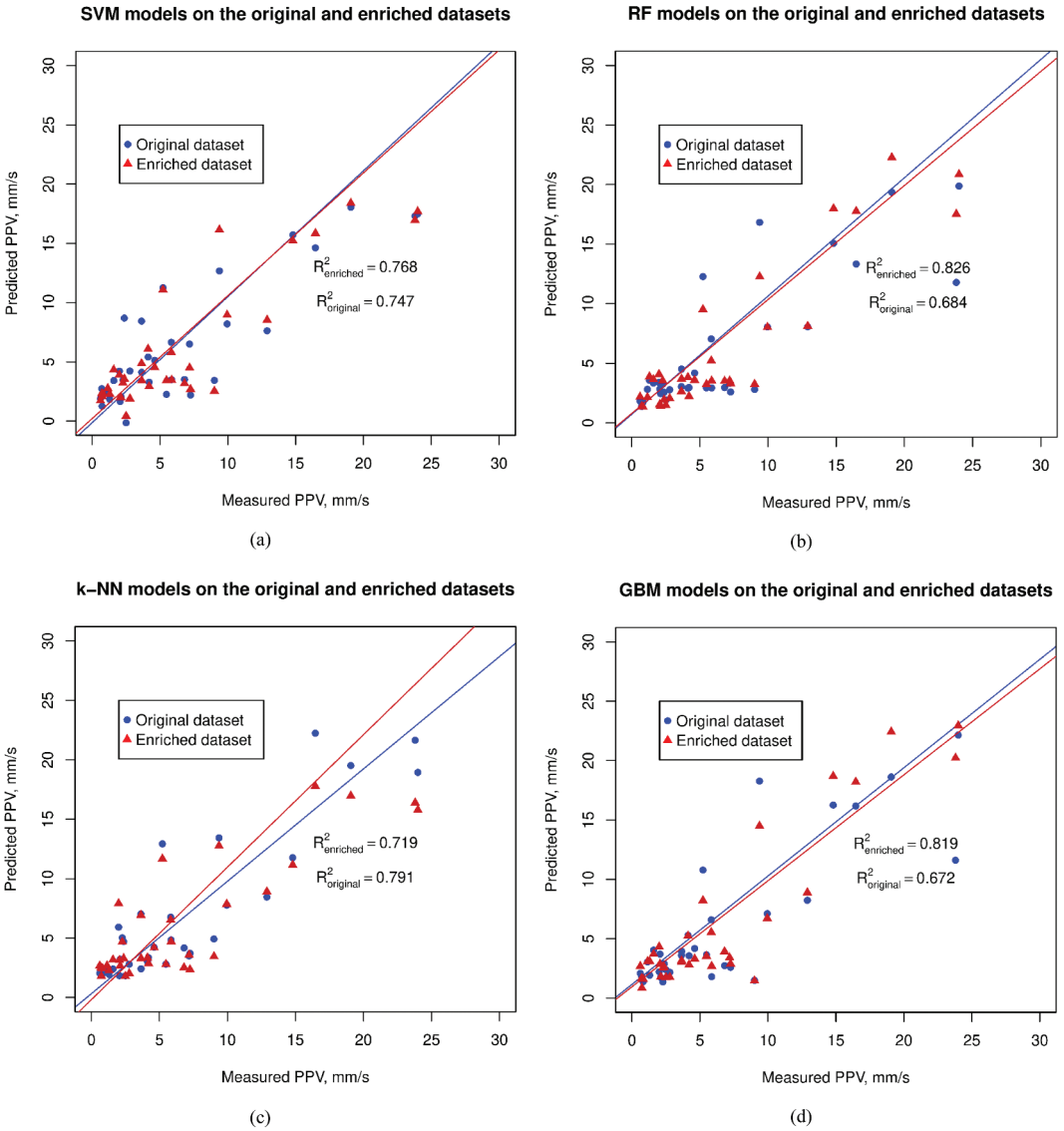


Figure 18. Regression analysis between the measured and predicted PPVs by the developed models on the original and enriched datasets (a) SVM models; (b) RF models; (c) k-NN models; (d) GBM models.

the model but less influential than ‘q’ with values of 27.74636 and 24.41886 respectively. ‘a’ has moderate importance value (12.91294) so it has moderate impact on the model’s predictions. Finally ‘L’ and ‘S’ have lower importance values than others so they have less influence on the model’s decisions (12.7016 and 10.55675 respectively).

5. Conclusions and future work

Based on the results presented, this study proposed a novel feature enrichment approach using the Cubist algorithm to enhance the accuracy of machine learning models in predicting blast-induced ground vibration in open-pit mines. The procedure involved introducing an additional input variable generated from the Cubist model, and the enriched dataset significantly improved the

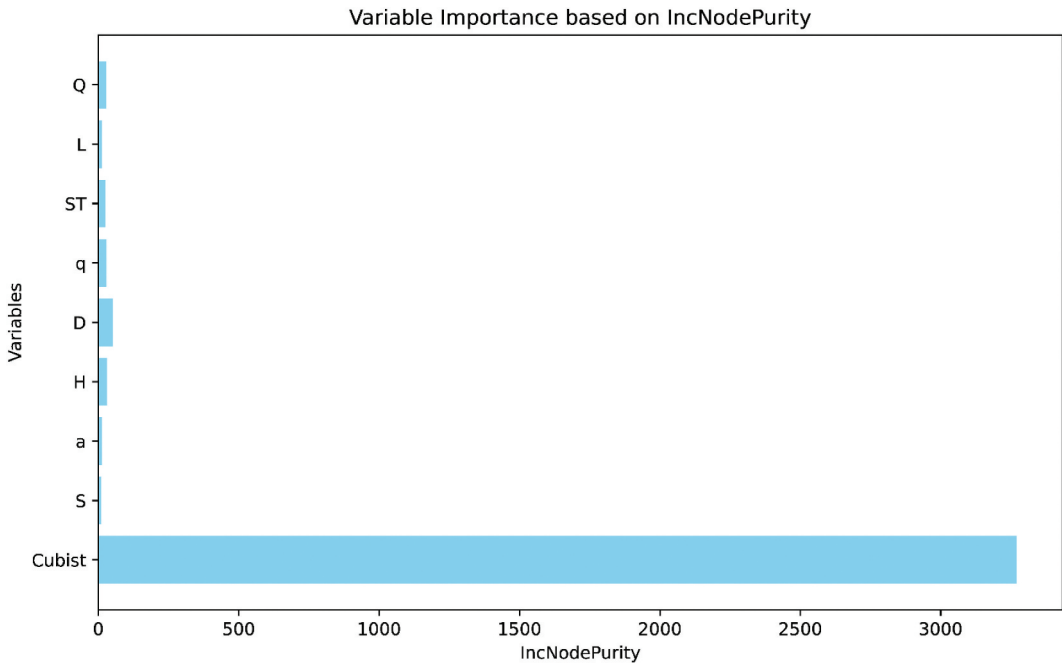


Figure 19. Variable importance analysis of the best predictive model.

performance of multiple machine learning models including SVM, RF, k-NN, and GBM. Among these models, the Random Forest model applied on the enriched dataset achieved the highest accuracy with an RMSE of 2.611, R^2 of 0.826, and MAE of 2.083. The results indicate that the RF model performed exceptionally well with the enriched dataset. Boxplots of residuals and residual trend analysis confirmed that enrichment reduced model bias and variance while improving generalisation capability. Therefore, the feature enrichment technique using Cubist algorithm was the most effective approach among the methods used in this study.

However, closer inspection of the study reveals some limitations that need to be addressed. First, the dataset used was derived from a single mining site, which may limit the generalisability of the models to other geological and operational conditions. This shows that the relevance of data with single-site approaches might be lower than with multi-site techniques in some cases. Second, while the Cubist algorithm effectively enriched the dataset in this case, the approach may require adaptation or validation in scenarios with significantly different data structures or distributions. The building process of the enrichment technique was more straightforward than traditional feature engineering methods. Finally, the study focused only on model accuracy; real-time deployability, computational cost, and interpretability were not addressed. Accordingly, these limitations were significant to guarantee the practical applicability of the proposed approach.

To improve upon these limitations, future work could explore several directions based on the findings of this study. First, integrating datasets from multiple mines with diverse geological conditions could improve the robustness and transferability of the models. The results show that multi-site validation, in addition to single-site testing, would be an influential factor for blast-induced ground vibration prediction in practical engineering. Second, hybrid feature enrichment strategies combining multiple ML-based transformations or physics-informed features may further enhance performance. However, the building process of hybrid approaches would be more complicated than that of the single Cubist-based method. Additionally, future studies should evaluate the real-time applicability of such models, incorporating continuous sensor data streams and edge

computing solutions. The performance of real-time models might be slightly different from that of offline models. Finally, explainable AI techniques could be employed to provide deeper insights into how enriched variables affect model predictions, supporting decision-making in practical blasting operations. Therefore, a procedure for analysing the degree of influence of enriched variables should be performed to identify the significant factors affecting the quality of the model.

In conclusion, this study highlights the value of ML-based feature enrichment as a complementary technique to model optimisation, offering a promising direction for enhancing prediction reliability in blast-induced ground vibration and broader geotechnical applications. The results indicate that the proposed approach performed very well in improving model accuracy with significant enhancement in RMSE, R^2 , and MAE values on the testing dataset. However, future research should address the identified limitations to ensure broader applicability and practical deployment of the enrichment methodology.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Hoang Nguyen  <http://orcid.org/0000-0001-6122-8314>

References

- [1] J. Zhou, Y. Lu, Q. Tian, H. Liu, M. Hasanipanah, and J. Huang, *Advanced machine learning methods for prediction of blast-induced flyrock using hybrid SVR methods*, CMES-Comput. Modeling Eng. Sci. 140 (2) (2024), pp. 1595–1617. doi:[10.32604/cmesci.2024.048398](https://doi.org/10.32604/cmesci.2024.048398).
- [2] X. Ding, M. Hasanipanah, and D.V. Ulrikh, *Hybrid metaheuristic optimization algorithms with least-squares support vector machine and boosted regression tree models for prediction of air-blast due to mine blasting*, Nat. Resour. Res. 33 (3) (2024), pp. 1349–1363. doi:[10.1007/s11053-024-10329-1](https://doi.org/10.1007/s11053-024-10329-1).
- [3] H. Nguyen, X.-N. Bui, and E. Topal, *Enhancing predictions of blast-induced ground vibration in open-pit mines: Comparing swarm-based optimization algorithms to optimize self-organizing neural networks*, Int. J. Coal Geol. 275 (2023), p. 104294. doi:[10.1016/j.coal.2023.104294](https://doi.org/10.1016/j.coal.2023.104294).
- [4] T. Singh, L. Dontha, and V. Bhardwaj, *Study into blast vibration and frequency using ANFIS and MVRA*, Min. Tech. 117 (3) (2008), pp. 116–121. doi:[10.1179/037178409X405741](https://doi.org/10.1179/037178409X405741).
- [5] M. Khandelwal and T. Singh, *Prediction of blast-induced ground vibration using artificial neural network*, Int. J. Rock Mech. Min. Sci. 46 (7) (2009), pp. 1214–1222. doi:[10.1016/j.ijrmms.2009.03.004](https://doi.org/10.1016/j.ijrmms.2009.03.004).
- [6] M. Monjezi, M. Ahmadi, M. Sheikhan, A. Bahrami, and A.R. Salimi, *Predicting blast-induced ground vibration using various types of neural networks*, Soil Dyn. Earthquake Eng. 30 (11) (2010), pp. 1233–1236. doi:[10.1016/j.soildyn.2010.05.005](https://doi.org/10.1016/j.soildyn.2010.05.005).
- [7] M. Khandelwal, *Evaluation and prediction of blast-induced ground vibration using support vector machine*, Int. J. Rock Mech. Min. Sci. 47 (3) (2010), pp. 509–516. doi:[10.1016/j.ijrmms.2010.01.007](https://doi.org/10.1016/j.ijrmms.2010.01.007).
- [8] A. Verma and T. Singh, *Intelligent systems for ground vibration measurement: A comparative study*, Eng. Comput. 27 (3) (2011), pp. 225–233. doi:[10.1007/s00366-010-0193-7](https://doi.org/10.1007/s00366-010-0193-7).
- [9] D.J. Armaghani, M. Hajihassani, E.T. Mohamad, A. Marto, and S.A. Noorani, *Blasting-induced flyrock and ground vibration prediction through an expert artificial neural network based on particle swarm optimization*, Arab. J. Geo. Sci. 7 (12) (2014), pp. 5383–5396. doi:[10.1007/s12517-013-1174-0](https://doi.org/10.1007/s12517-013-1174-0).
- [10] M. Hajihassani, D. Jahed Armaghani, A. Marto, and E. Tonnizam Mohamad, *Ground vibration prediction in quarry blasting through an artificial neural network optimized by imperialist competitive algorithm*, Bull. Eng. Geol. Environ. 74 (3) (2015), pp. 873–886. doi:[10.1007/s10064-014-0657-x](https://doi.org/10.1007/s10064-014-0657-x).
- [11] S.R. Dindarloo, *Prediction of blast-induced ground vibrations via genetic programming*, Int. J. Min. Sci. Technol. 25 (6) (2015), pp. 1011–1015. doi:[10.1016/j.ijmst.2015.09.020](https://doi.org/10.1016/j.ijmst.2015.09.020).
- [12] M. Saadat, A. Hasanzade, and M. Khandelwal, *Differential evolution algorithm for predicting blast induced ground vibrations*, Int. J. Rock Mech. Min. Sci. 77 (2015), pp. 97–104. doi:[10.1016/j.ijrmms.2015.03.020](https://doi.org/10.1016/j.ijrmms.2015.03.020).
- [13] M. Amiri, H. Bakhshandeh Amnieh, M. Hasanipanah, and L. Mohammad Khanli, *A new combination of artificial neural network and K-nearest neighbors models to predict blast-induced ground vibration and air-overpressure*, Eng. Comput. 32 (4) (2016), pp. 631–644. doi:[10.1007/s00366-016-0442-5](https://doi.org/10.1007/s00366-016-0442-5).

- [14] J. Zhou, X. Shi, and X. Li, *Utilizing gradient boosted machine for the prediction of damage to residential structures owing to blasting vibrations of open pit mining*, J. Vib. Control 22 (19) (2016), pp. 3986–3997. doi:10.1177/1077546314568172.
- [15] R.S. Faradonbeh and M. Monjezi, *Prediction and minimization of blast-induced ground vibration using two robust meta-heuristic algorithms*, Eng. Comput. 33 (4) (2017), pp. 835–851. doi:10.1007/s00366-017-0501-6.
- [16] A. Shahnazar, H. Nikafshan Rad, M. Hasanipanah, M.M. Tahir, D.M.M. Jahed Armaghani, and M. Ghorogi, *A new developed approach for the prediction of ground vibration using a hybrid PSO-optimized ANFIS-based model*, Environ. Earth Sci. 76 (15) (2017), p. 527. doi:10.1007/s12665-017-6864-6.
- [17] Hasanipanah, R.S. Faradonbeh, H.B. Amnieh, D.J. Armaghani, and M. Monjezi, *Forecasting blast-induced ground vibration developing a CART model*, Eng. Comput. 33 (2) (2017), pp. 307–316. doi:10.1007/s00366-016-0475-9.
- [18] K. Taheri, M. Hasanipanah, S.B. Golzar, and M.Z. Majid, *A hybrid artificial bee colony algorithm-artificial neural network for forecasting the blast-produced ground vibration*, Eng. Comput. 33 (3) (2017), pp. 689–700. doi:10.1007/s00366-016-0497-3.
- [19] D.J. Armaghani, M. Hasanipanah, H.B. Amnieh, and E.T. Mohamad, *Feasibility of ICA in approximating ground vibration resulting from mine blasting*, Neural Comput. Appl. 29 (9) (2018), pp. 457–465. doi:10.1007/s00521-016-2577-0.
- [20] H. Sheykhi, R. Bagherpour, E. Ghasemi, and H. Kalhori, *Forecasting ground vibration due to rock blasting: A hybrid intelligent approach using support vector regression and fuzzy C-means clustering*, Eng. Comput. 34 (2) (2018), pp. 357–365. doi:10.1007/s00366-017-0546-6.
- [21] Hasanipanah, H. Bakhshandeh Amnieh, H. Khamesi, D. Jahed Armaghani, S. Bagheri Golzar, and A. Shahnazar, *Prediction of an environmental issue of mine blasting: An imperialistic competitive algorithm-based fuzzy system*, Int. J. Environ. Sci. Technol. 15 (3) (2018), pp. 551–560. doi:10.1007/s13762-017-1395-y.
- [22] Y. Azimi, S.H. Khoshrou, and M. Osanloo, *Prediction of blast induced ground vibration (BIGV) of quarry mining using hybrid genetic algorithm optimized artificial neural network*, Measurement 147 (2019), p. 106874. doi:10.1016/j.measurement.2019.106874.
- [23] H. Nguyen, Y. Choi, X.N. Bui, and T. Nguyen-Thoi, *Predicting blast-induced ground vibration in open-pit mines using vibration sensors and support vector regression-based optimization algorithms*, Sensors 20 (1) (2019), p. 132. doi:10.3390/s20010132.
- [24] X.-N. Bui, P. Jaroonpattanapong, H. Nguyen, Q.H. Tran, and N.Q. Long, *A novel hybrid model for predicting blast-induced ground vibration based on k-nearest neighbors and particle swarm optimization*, Sci. Rep. 9 (1) (2019), pp. 1–14. doi:10.1038/s41598-019-50262-5.
- [25] H. Nguyen, X.N. Bui, Q.H. Tran, and N.L. Mai, *A new soft computing model for estimating and controlling blast-produced ground vibration based on hierarchical K-means clustering and Cubist algorithms*, Appl. Soft Comput. 77 (2019), pp. 376–386. doi:10.1016/j.asoc.2019.01.042.
- [26] S.A. Hosseini, A. Tavana, S.M. Abdolahi, and S. Darvishmaslak, *Prediction of blast-induced ground vibrations in quarry sites: A comparison of GP, RSM and MARS*, Soil Dyn. Earthquake Eng. 119 (2019), pp. 118–129. doi:10.1016/j.soildyn.2019.01.011.
- [27] H. Nguyen, X.N. Bui, H.B. Bui, and D.T. Cuong, *Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: A case study*, Acta geophys. 67 (2) (2019), pp. 477–490. doi:10.1007/s11600-019-00268-4.
- [28] H. Nguyen, X.N. Bui, Q.H. Tran, and H. Moayedi, *Predicting blast-induced peak particle velocity using BGAMs, ANN and SVM: A case study at the nui beo open-pit coal mine in Vietnam*, Environ. Earth Sci. 78 (15) (2019), p. 479. doi:10.1007/s12665-019-8491-x.
- [29] H. Yang, M. Hasanipanah, M.M. Tahir, and D.T. Bui, *Intelligent prediction of blasting-induced ground vibration using ANFIS optimized by GA and PSO*, Nat. Resour. Res. 29 (2) (2019), pp. 739–750. doi:10.1007/s11053-019-09515-3.
- [30] M. Amiri, M. Hasanipanah, and H. Bakhshandeh Amnieh, *Predicting ground vibration induced by rock blasting using a novel hybrid of neural network and itemset mining*, Neural Comput. Appl. 32 (18) (2020), pp. 14681–14699. doi:10.1007/s00521-020-04822-w.
- [31] Y. Shang, H. Nguyen, X.N. Bui, Q.H. Tran, and H. Moayedi, *A novel artificial intelligence approach to predict blast-induced ground vibration in open-pit mines based on the firefly algorithm and artificial neural network*, Nat. Resour. Res. 29 (2) (2019), pp. 723–737. doi:10.1007/s11053-019-09503-7.
- [32] J. Zhou, P.G. Asteris, D.J. Armaghani, and B.T. Pham, *Prediction of ground vibration induced by blasting operations through the use of the Bayesian network and random forest models*, Soil Dyn. Earthquake Eng. 139 (2020), pp. 106390. doi:10.1016/j.soildyn.2020.106390.
- [33] H. Nguyen, C. Drebenstedt, X.N. Bui, and D.T. Bui, *Prediction of blast-induced ground vibration in an open-pit mine by a novel hybrid model based on clustering and artificial neural network*, Nat Resour Res. 29 (2) (2020), pp. 691–709. doi:10.1007/s11053-019-09470-z.

- [34] G. Li, D. Kumar, P. Samui, R.H. Nikafshan, B. Roy, and M. Hasanipanah, *Developing a new computational intelligence approach for approximating the blast-induced ground vibration*, Appl. Sci. 10 (2) (2020), p. 434. doi:10.3390/app10020434.
- [35] X. Zhang, H. Nguyen, X.N. Bui, Q.H. Tran, D.A. Nguyen, D.T. Bui, and H. Moayedi, *Novel soft computing model for predicting blast-induced ground vibration in open-pit mines based on particle swarm optimization and XGBoost*, Nat Resour Res. 29 (2) (2020), pp. 711–721. doi:10.1007/s11053-019-09492-7.
- [36] H. Yang, H. Nikafshan Rad, M. Hasanipanah, H. Bakhshandeh Amnieh, and A. Nekouie, *Prediction of vibration velocity generated in mine blasting using support vector regression improved by optimization algorithms*, Nat Resour Res. 29 (2) (2020), pp. 807–830. doi:10.1007/s11053-019-09597-z.
- [37] X.-N. Bui, Y. Choi, V. Atrushkevich, H. Nguyen, and Q.H. Tran, *Prediction of blast-induced ground vibration intensity in open-pit mines using unmanned aerial vehicle and a novel intelligence system*, Nat Resour Res. 29 (2) (2020), pp. 771–790. doi:10.1007/s11053-019-09573-7.
- [38] Z. Yu, X. Shi, J. Zhou, X. Chen, and X. Qiu, *Effective assessment of blast-induced ground vibration using an optimized random forest model based on a Harris Hawks optimization algorithm*, Appl. Sci. 10 (4) (2020), p. 1403. doi:10.3390/app10041403.
- [39] Z. Yu, X. Shi, J. Zhou, Y. Gou, X. Huo, J. Zhang, and D.J. Armaghani, *A new multikernel relevance vector machine based on the HPSOGWO algorithm for predicting and controlling blast-induced ground vibration*, Eng. With Comput. 38 (2) (2022), pp. 1905–1920. doi:10.1007/s00366-020-01136-2.
- [40] Z. Ding, H. Nguyen, X.N. Bui, J. Zhou, and H. Moayedi, *Computational intelligence model for estimating intensity of blast-induced ground vibration in a mine based on imperialist competitive and extreme gradient boosting algorithms*, Nat Resour Res. 29 (2) (2020), pp. 751–769. doi:10.1007/s11053-019-09548-8.
- [41] C.K. Arthur, V.A. Temeng, and Y.Y. Ziggah, *A self-adaptive differential evolutionary extreme learning machine (SaDE-ELM): A novel approach to blast-induced ground vibration prediction*, SN Appl Sci. 2 (11) (2020), pp. 1845. doi:10.1007/s42452-020-03611-3.
- [42] W. Zhu, H. Nikafshan Rad, and M. Hasanipanah, *A chaos recurrent ANFIS optimized by PSO to predict ground vibration generated in rock blasting*, Appl. Soft Comput. 108 (2021), p. 107434. doi:10.1016/j.asoc.2021.107434.
- [43] H. Nguyen and X.-N. Bui, *A novel hunger games search optimization-based artificial neural network for predicting ground vibration intensity induced by mine blasting*, Nat Resour Res. 30 (5) (2021), pp. 3865–3880. doi:10.1007/s11053-021-09903-8.
- [44] X.-N. Bui, H. Nguyen, Q.H. Tran, D.A. Nguyen, and H.B. Bui, *Predicting ground vibrations due to mine blasting using a novel artificial neural network-based cuckoo search optimization*, Nat Resour Res. 30 (3) (2021), pp. 2663–2685. doi:10.1007/s11053-021-09823-7.
- [45] H. Fattahi and M. Hasanipanah, *Prediction of blast-induced ground vibration in a mine using relevance vector regression optimized by metaheuristic algorithms*, Nat Resour Res. 30 (2) (2021), pp. 1849–1863. doi:10.1007/s11053-020-09764-7.
- [46] A.I. Lawal, S. Kwon, O.S. Hammed, and M.A. Idris, *Blast-induced ground vibration prediction in granite quarries: An application of gene expression programming, ANFIS, and sine cosine algorithm optimized ANN*, Int. J. Min. Sci. Technol. 31 (2) (2021), pp. 265–277. doi:10.1016/j.ijmst.2021.01.007.
- [47] J. Zhou, Y. Qiu, M. Khandelwal, S. Zhu, and X. Zhang, *Developing a hybrid model of Jaya algorithm-based extreme gradient boosting machine to estimate blast-induced ground vibrations*, Int. J. Rock Mech. Min. Sci. 145 (2021), pp. 104856. doi:10.1016/j.ijrmms.2021.104856.
- [48] D. Jahed Armaghani, D. Kumar, P. Samui, M. Hasanipanah, and B. Roy, *A novel approach for forecasting of ground vibrations resulting from blasting: Modified particle swarm optimization coupled extreme learning machine*, Eng. Comput. 37 (4) (2021), pp. 3221–3235. doi:10.1007/s00366-020-00997-x.
- [49] B. Ke, H. Nguyen, X.N. Bui, and R. Costache, *Estimation of ground vibration intensity induced by mine blasting using a state-of-the-art hybrid autoencoder neural network and support vector regression model*, Nat Resour Res. 30 (5) (2021), pp. 3853–3864. doi:10.1007/s11053-021-09890-w.
- [50] X. Ding, M. Hasanipanah, H. Nikafshan Rad, and W. Zhou, *Predicting the blast-induced vibration velocity using a bagged support vector regression optimized with firefly algorithm*, Eng. Comput. 37 (3) (2021), pp. 2273–2284. doi:10.1007/s00366-020-00937-9.
- [51] H. Nguyen, X.N. Bui, Q.H. Tran, D.A. Nguyen, L.T. Hoa, Q.T. Le, and L.T. Giang, *Predicting blast-induced ground vibration in open-pit mines using different nature-inspired optimization algorithms and deep neural network*, Nat Resour Res. 30 (6) (2021), pp. 4695–4717. doi:10.1007/s11053-021-09896-4.
- [52] A.I. Lawal, S. Kwon, and G.Y. Kim, *Prediction of the blast-induced ground vibration in tunnel blasting using ANN, moth-flame optimized ANN, and gene expression programming*, Acta geophys. 69 (1) (2021), pp. 161–174. doi:10.1007/s11600-020-00532-y.
- [53] X.-N. Bui, H. Nguyen, Q.H. Tran, D.A. Nguyen, and H.B. Bui, *Predicting blast-induced ground vibration in quarries using adaptive fuzzy inference neural network and moth-flame optimization*, Nat Resour Res. 30 (6) (2021), pp. 4719–4734. doi:10.1007/s11053-021-09968-5.

- [54] Y. Qiu, J. Zhou, M. Khandelwal, H. Yang, P. Yang, and C. Li, *Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration*, Eng. Comput. 38 (5) (2022), pp. 4145–4162. doi:10.1007/s00366-021-01393-9.
- [55] A. Abbaszadeh Shahri, F. Pashamohammadi, R. Asheghi, and H. Abbaszadeh Shahri, *Automated intelligent hybrid computing schemes to predict blasting induced ground vibration*, Eng. Comput. 38 (4) (2022), pp. 3335–3349. doi:10.1007/s00366-021-01444-1.
- [56] H. Nguyen, X.N. Bui, Q.H. Tran, H.A. Nguyen, D.A. Nguyen, L.T. Hoa, and Q.T. Le, *Prediction of ground vibration intensity in mine blasting using the novel hybrid MARS-PSO-MLP model*, Eng. Comput. 38 (5) (2022), pp. 4007–4025. doi:10.1007/s00366-021-01332-8.
- [57] H. Nguyen and X.-N. Bui, *Optimized adaptive neuro-fuzzy inference system for predicting blast-induced ground vibration in quarries based on hunger games search optimization*, Int. J. Min. Reclam. Environ. 36 (10) (2022), pp. 724–748. doi:10.1080/17480930.2022.2131137.
- [58] H. Nguyen, X.-N. Bui, and E. Topal, *Reliability and availability artificial intelligence models for predicting blast-induced ground vibration intensity in open-pit mines to ensure the safety of the surroundings*, Reliab. Eng. Syst. Saf. 231 (2023), pp. 109032. doi:10.1016/j.res.2022.109032.
- [59] A.G. Yardimci and M. Erkayaoglu, *Simulation of blast-induced ground vibrations using a machine learning-assisted Mechanical framework*, Environ. Earth Sci. 82 (21) (2023), pp. 508. doi:10.1007/s12665-023-11194-6.
- [60] H. Nguyen, Y. Choi, M. Monjezi, N. Van Thieu, and T.T. Tran, *Predicting different components of blast-induced ground vibration using earthworm optimisation-based adaptive neuro-fuzzy inference system*, Int. J. Min. Reclam. Environ. 38(2) (2024) pp. 1–28.
- [61] S. Alzabeebee, M. Jamei, M. Hasanipanah, H.B. Amnieh, M. Karbasi, and S. Keawsawasvong, *Development of a new explicit soft computing model to predict the blast-induced ground vibration*, Geomech. Eng. 30 (6) (2022), pp. 551–564.
- [62] X. Ding, M. Bahadori, M. Hasanipanah, and R.A. Abdullah, *Predicting the rock fragmentation in surface mines using optimized radial basis function and cascaded forward neural network models*, Geomech. Eng. 33 (6) (2023), pp. 567–581.
- [63] K. Peng, J. Zeng, D.J. Armaghani, M. Hasanipanah, and Q. Chen, *A novel combination of gradient boosted tree and optimized ANN models for forecasting ground vibration due to quarry blasting*, Nat Resour Res. 30 (6) (2021), pp. 4657–4671. doi:10.1007/s11053-021-09899-1.
- [64] Rulequest. *Data Mining with Cubist*, (2016).
- [65] R. Quinlan, *Data Mining Tools See5 and C5. 0*, 2004.
- [66] Rulequest, *Data mining with cubist*, RuleQuest Research Pty Ltd., St. Ives, NSW, Australia, 2016; available at <https://www.rulequest.com/cubist-info.html>.
- [67] C. Cortes and V. Vapnik, *Support vector machine*, Mach Learn. 20 (3) (1995), pp. 273–297. doi:10.1023/A:1022627411411.
- [68] N.M. Shahani and Z. Ye, *Application of gradient boosting machine learning algorithms to predict uniaxial compressive strength of soft sedimentary rocks at Thar Coalfield*, Adv. Civ. Eng. 2021 (1) (2021), p. 2565488. doi:10.1155/2021/2565488.
- [69] R. Zhang, Y. Li, G. At, W. Zhang, and Z. Chen, *Analysis of ground surface settlement in anisotropic clays using extreme gradient boosting and random forest regression models*, J. Rock Mech. Geotech. Eng. 13 (6) (2021), pp. 1478–1484. doi:10.1016/j.jrmge.2021.08.001.
- [70] J. Zhou, E. Li, S. Yang, M. Wang, X. Shi, S. Yao, and H.S. Mitri, *Slope stability prediction for circular mode failure using gradient boosting machine approach based on an updated database of case histories*, Saf. Sci. 118 (2019), pp. 505–518. doi:10.1016/j.ssci.2019.05.046.
- [71] M. Ahmad, A. Murari, K.R. Kashyzadeh, S. Keawsawasvong, M.M. Sabri, I. Jamil, and A.C. Alguno, *Extreme gradient boosting algorithm for predicting shear strengths of rockfill materials*, Complexity 2022 (1) (2022), p. 9415863. doi:10.1155/2022/9415863.
- [72] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, *Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization*, Geosci. Front. 12 (1) (2021), pp. 469–477. doi:10.1016/j.gsf.2020.03.007.
- [73] S. Ray, R. Mm, M. Haque, H. Mw, and M.M. Alam, *Performance evaluation of SVM and GBM in predicting compressive and splitting tensile strength of concrete prepared with ceramic waste and nylon fiber*, J. Educ. Chang. King Saud Univ.-Eng. Sci. 35 (2) (2023), pp. 92–100. doi:10.1016/j.jksues.2021.02.009.
- [74] S. Lin, H. Zheng, B. Han, Y. Li, C. Han, and W. Li, *Comparative performance of eight ensemble learning approaches for the development of models of slope stability prediction*, Acta GeoTech. 17 (4) (2022), pp. 1477–1502. doi:10.1007/s11440-021-01440-1.
- [75] M. Protodiakonov, M. Koifman, S. Chirkov, M. Kuntish, and R. Tedder, *Rock Strength Passports and Methods for Their Determination*, Moscow, Russia: Nauka, 1964.

APPENDIX

Table A1. Training performance of the GBM model on the original dataset.

No.	shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	R ²	MAE
1	0.001149	4	22	2351	NA	NA	NA
2	0.014694	4	13	3674	4.386575	0.596339	3.276882
3	0.015137	2	9	1375	4.357478	0.614967	3.103433
4	0.022373	7	21	1706	NA	NA	NA
5	0.025581	10	25	2057	NA	NA	NA
6	0.038364	9	5	270	3.807603	0.721686	2.642149
7	0.043673	7	17	2316	4.501535	0.597116	3.436295
8	0.046366	3	25	3254	NA	NA	NA
9	0.052673	8	10	302	4.338114	0.61793	3.104367
10	0.055072	1	7	1244	4.557777	0.596935	3.25165
11	0.062186	4	16	1354	4.449098	0.598358	3.360239
12	0.07301	2	16	2593	4.862057	0.525175	3.721051
13	0.073786	9	14	2881	4.818427	0.52197	3.626621
14	0.089567	10	10	2542	5.04133	0.490433	3.727854
15	0.089855	5	25	1334	NA	NA	NA
16	0.092956	8	9	4791	4.741753	0.538116	3.480194
17	0.094126	2	13	1952	4.817365	0.514671	3.624075
18	0.094211	4	6	3160	3.861581	0.702569	2.703345
19	0.096141	1	10	2614	4.974564	0.487549	3.796079
20	0.097892	9	7	4900	4.043923	0.669902	2.832699
21	0.10093	7	18	4679	5.346499	0.466774	3.997229
22	0.103078	5	8	3989	4.338295	0.608299	3.140508
23	0.108618	4	10	3436	5.126263	0.469406	3.838623
24	0.11321	3	12	2849	5.210686	0.465968	3.947198
25	0.114249	9	23	109	NA	NA	NA
26	0.119182	1	15	4653	5.399675	0.433995	4.058753
27	0.122534	9	11	2699	5.235875	0.445397	3.906242
28	0.139556	9	15	4415	5.370103	0.441588	4.036906
29	0.141093	9	12	1059	5.012667	0.495794	3.701935
30	0.146744	3	7	1855	4.205214	0.636156	2.999669
31	0.15009	5	7	1395	4.131817	0.644681	2.964977
32	0.159028	9	24	1964	NA	NA	NA
33	0.15966	1	21	1522	NA	NA	NA
34	0.165487	6	15	417	4.426842	0.600403	3.331322
35	0.176697	6	12	4778	5.480275	0.417829	4.197235
36	0.183148	2	21	2283	NA	NA	NA
37	0.196527	8	14	2822	5.261128	0.449481	4.014014
38	0.216703	5	20	558	NA	NA	NA
39	0.224594	9	20	3704	NA	NA	NA
40	0.230323	6	12	3120	5.523926	0.413666	4.230654
41	0.232531	6	16	4738	5.772441	0.388903	4.297873
42	0.233319	6	22	4998	NA	NA	NA
43	0.24134	7	23	3803	NA	NA	NA
44	0.245642	1	15	4995	5.777924	0.378315	4.388428
45	0.247982	9	21	3633	NA	NA	NA
46	0.248494	1	15	2546	5.497142	0.416081	4.076564
47	0.260614	5	20	2265	NA	NA	NA
48	0.262785	10	18	3957	5.931029	0.403078	4.311614

(Continued)

Table A1. (Continued).

No.	shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	R ²	MAE
49	0.267608	2	23	3196	NA	NA	NA
50	0.279091	4	13	3285	5.551709	0.414331	4.323517
51	0.279216	7	22	1679	NA	NA	NA
52	0.279666	9	22	1881	NA	NA	NA
53	0.299059	6	13	2556	5.589536	0.404378	4.284632
54	0.30877	3	7	4937	4.196216	0.658856	2.996316
55	0.308929	5	7	654	4.276112	0.63844	3.131448
56	0.31792	3	20	1092	NA	NA	NA
57	0.32051	9	24	3508	NA	NA	NA
58	0.321639	5	13	2715	5.672021	0.405108	4.370073
59	0.331134	1	20	1631	NA	NA	NA
60	0.339531	7	9	3835	4.790573	0.550508	3.614883
61	0.341123	10	15	479	4.746272	0.539009	3.594269
62	0.343292	8	14	2166	5.485249	0.416689	4.180877
63	0.349978	7	7	3301	4.399558	0.609307	3.177704
64	0.372538	8	24	3930	NA	NA	NA
65	0.390364	6	11	3858	5.525506	0.415499	4.294375
66	0.392513	1	15	4066	5.90378	0.398972	4.386474
67	0.39682	3	6	4504	4.446193	0.630932	3.101051
68	0.400283	1	12	2241	5.734532	0.404337	4.388661
69	0.405764	8	7	2119	4.410707	0.625047	3.196358
70	0.406992	2	11	2043	5.79623	0.39366	4.505007
71	0.411487	4	18	4462	6.267665	0.344252	4.497325
72	0.414081	8	24	4688	NA	NA	NA
73	0.415132	9	21	8	NA	NA	NA
74	0.429067	3	20	596	NA	NA	NA
75	0.430256	5	6	1665	4.148512	0.638058	3.061916
76	0.433648	2	5	336	4.133967	0.692351	2.95568
77	0.438964	8	22	758	NA	NA	NA
78	0.444245	1	19	910	NA	NA	NA
79	0.461566	5	15	2254	5.909629	0.370731	4.389246
80	0.479566	6	16	4509	6.292796	0.333832	4.704531
81	0.48372	5	19	2724	NA	NA	NA
82	0.486818	5	12	3375	5.832111	0.39289	4.591485
83	0.487603	9	14	1314	5.557005	0.419518	4.164251
84	0.496308	7	21	1293	NA	NA	NA
85	0.500043	5	11	569	5.504494	0.43363	4.246817
86	0.503435	6	23	4489	NA	NA	NA
87	0.506523	2	5	4456	4.29103	0.650071	3.147511
88	0.507874	10	8	3106	4.762341	0.566174	3.487954
89	0.508395	3	22	3442	NA	NA	NA
90	0.521118	5	24	3384	NA	NA	NA
91	0.522938	4	21	3465	NA	NA	NA
92	0.543435	9	18	2092	6.130737	0.380738	4.420247
93	0.550859	4	15	522	5.318079	0.446274	4.07014
94	0.551176	4	6	38	4.80383	0.562812	3.51144
95	0.555793	1	7	1248	5.212275	0.515183	3.82213
96	0.558888	6	9	2905	5.448393	0.425866	4.084177
97	0.572459	7	20	542	NA	NA	NA
98	0.586788	3	6	141	4.562059	0.650487	3.267443
99	0.595799	9	12	3018	6.102514	0.37856	4.737745
100	0.596342	1	8	625	5.451697	0.462033	4.084655

Note: NA is not a number.

Table A2. Training performance of the GBM model on the enriched dataset.

No.	shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	R ²	MAE
1	0.001149	4	22	2351	NA	NA	NA
2	0.014694	4	13	3674	4.132822	0.676041	2.964837
3	0.015137	2	9	1375	3.750214	0.738641	2.593016
4	0.022373	7	21	1706	NA	NA	NA
5	0.025581	10	25	2057	NA	NA	NA
6	0.038364	9	5	270	2.907844	0.835758	2.058606
7	0.043673	7	17	2316	4.39439	0.621619	3.349479
8	0.046366	3	25	3254	NA	NA	NA
9	0.052673	8	10	302	3.83788	0.732117	2.677125
10	0.055072	1	7	1244	3.407951	0.771625	2.409861
11	0.062186	4	16	1354	4.305456	0.630826	3.243477
12	0.07301	2	16	2593	4.608934	0.577006	3.495977
13	0.073786	9	14	2881	4.469352	0.603151	3.28502
14	0.089567	10	10	2542	4.335857	0.62362	3.191693
15	0.089855	5	25	1334	NA	NA	NA
16	0.092956	8	9	4791	4.133456	0.654445	2.988416
17	0.094126	2	13	1952	4.543653	0.58899	3.340356
18	0.094211	4	6	3160	3.258323	0.783726	2.309573
19	0.096141	1	10	2614	4.521952	0.613091	3.397768
20	0.097892	9	7	4900	3.634245	0.735493	2.526311
21	0.10093	7	18	4679	5.276788	0.469044	4.04393
22	0.103078	5	8	3989	3.935165	0.68899	2.842592
23	0.108618	4	10	3436	4.463212	0.598969	3.299585
24	0.11321	3	12	2849	4.721845	0.559761	3.548362
25	0.114249	9	23	109	NA	NA	NA
26	0.119182	1	15	4653	4.959324	0.510106	3.774002
27	0.122534	9	11	2699	4.54165	0.591558	3.347956
28	0.139556	9	15	4415	5.004606	0.495628	3.82407
29	0.141093	9	12	1059	4.616532	0.577054	3.387119
30	0.146744	3	7	1855	3.548307	0.744323	2.508712
31	0.15009	5	7	1395	3.666741	0.733769	2.596106
32	0.159028	9	24	1964	NA	NA	NA
33	0.15966	1	21	1522	NA	NA	NA
34	0.165487	6	15	417	4.298111	0.640383	3.179265
35	0.176697	6	12	4778	4.857276	0.537211	3.650772
36	0.183148	2	21	2283	NA	NA	NA
37	0.196527	8	14	2822	4.927286	0.511634	3.707559
38	0.216703	5	20	558	NA	NA	NA
39	0.224594	9	20	3704	NA	NA	NA
40	0.230323	6	12	3120	4.784352	0.54934	3.60066
41	0.232531	6	16	4738	5.378636	0.462871	4.080597
42	0.233319	6	22	4998	NA	NA	NA
43	0.24134	7	23	3803	NA	NA	NA
44	0.245642	1	15	4995	5.423755	0.448184	4.145253
45	0.247982	9	21	3633	NA	NA	NA
46	0.248494	1	15	2546	5.204938	0.480468	3.958327
47	0.260614	5	20	2265	NA	NA	NA
48	0.262785	10	18	3957	5.953691	0.39355	4.499146
49	0.267608	2	23	3196	NA	NA	NA
50	0.279091	4	13	3285	4.99687	0.51441	3.862685

(Continued)

Table A2. (Continued).

No.	shrinkage	interaction.depth	n.minobsinnode	n.trees	RMSE	R ²	MAE
51	0.279216	7	22	1679	NA	NA	NA
52	0.279666	9	22	1881	NA	NA	NA
53	0.299059	6	13	2556	5.070293	0.50036	3.907939
54	0.30877	3	7	4937	3.903862	0.690117	2.812144
55	0.308929	5	7	654	3.761083	0.71131	2.673495
56	0.31792	3	20	1092	NA	NA	NA
57	0.32051	9	24	3508	NA	NA	NA
58	0.321639	5	13	2715	5.101958	0.508286	3.892647
59	0.331134	1	20	1631	NA	NA	NA
60	0.339531	7	9	3835	4.219115	0.659143	3.145494
61	0.341123	10	15	479	4.660949	0.581282	3.46629
62	0.343292	8	14	2166	5.119165	0.495268	3.939585
63	0.349978	7	7	3301	3.66555	0.718657	2.599368
64	0.372538	8	24	3930	NA	NA	NA
65	0.390364	6	11	3858	4.928496	0.498167	3.747441
66	0.392513	1	15	4066	5.550897	0.436401	4.240687
67	0.39682	3	6	4504	3.562806	0.739432	2.478925
68	0.400283	1	12	2241	4.978477	0.534831	3.852479
69	0.405764	8	7	2119	3.537495	0.735124	2.580363
70	0.406992	2	11	2043	4.818566	0.56994	3.625972
71	0.411487	4	18	4462	5.979894	0.352371	4.439227
72	0.414081	8	24	4688	NA	NA	NA
73	0.415132	9	21	8	NA	NA	NA
74	0.429067	3	20	596	NA	NA	NA
75	0.430256	5	6	1665	3.325489	0.786745	2.395864
76	0.433648	2	5	336	3.24987	0.775286	2.341304
77	0.438964	8	22	758	NA	NA	NA
78	0.444245	1	19	910	NA	NA	NA
79	0.461566	5	15	2254	5.556091	0.422588	4.30087
80	0.479566	6	16	4509	5.726127	0.407321	4.37984
81	0.48372	5	19	2724	NA	NA	NA
82	0.486818	5	12	3375	5.106695	0.499383	3.993079
83	0.487603	9	14	1314	5.19576	0.481449	3.950033
84	0.496308	7	21	1293	NA	NA	NA
85	0.500043	5	11	569	4.980263	0.550832	3.85729
86	0.503435	6	23	4489	NA	NA	NA
87	0.506523	2	5	4456	3.358763	0.778743	2.485004
88	0.507874	10	8	3106	4.118595	0.664599	2.899733
89	0.508395	3	22	3442	NA	NA	NA
90	0.521118	5	24	3384	NA	NA	NA
91	0.522938	4	21	3465	NA	NA	NA
92	0.543435	9	18	2092	5.868044	0.410356	4.409369
93	0.550859	4	15	522	5.013246	0.517255	3.79683
94	0.551176	4	6	38	3.525532	0.748744	2.688998
95	0.555793	1	7	1248	4.371677	0.630908	3.10677
96	0.558888	6	9	2905	4.62488	0.584359	3.368021
97	0.572459	7	20	542	NA	NA	NA
98	0.586788	3	6	141	3.939274	0.678588	2.870145
99	0.595799	9	12	3018	5.495522	0.499076	4.27109
100	0.596342	1	8	625	4.556901	0.604601	3.437141

Note: NA is not a number.